

Vol. IX Part I

PR/G2 -
May 1956

THE BRITISH JOURNAL OF STATISTICAL PSYCHOLOGY

EDITED BY
CYRIL BURT

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ALAN STUART
AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN	E. A. PEEL
M. S. BARTLETT	L. S. PENROSE
W. G. EMMETT	J. FRASER ROBERTS
M. G. KENDALL	A. RODGER
D. N. LAWLEY	W. STEPHENSON
E. S. PEARSON	P. E. VERNON

Managing Sub-editor J. W. WHITFIELD

Published by
THE BRITISH PSYCHOLOGICAL SOCIETY
Printed and distributed by
THE ABERDEEN UNIVERSITY PRESS LTD.
6 UPPER KIRKGATE
ABERDEEN

Price 20s. per part (U.S.A. \$3.50)

Subscription 30s. 6d. per volume (U.S.A. \$4.50)

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Statistical Psychology is issued by the British Psychological Society. The subscription price per volume is 30s. net for Members of the Society and 30s. 6d. (post free) for non-members. The subscription price in the U.S.A. is \$4.50 (post free). Members of the Society should send their subscriptions to THE SECRETARY, British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1: non-members to the ABERDEEN UNIVERSITY PRESS, 6, Upper Kirkgate, Aberdeen, Scotland. Until further notice the Journal will be issued in two parts each year, in May and November.

Papers for publication should be sent to Sir CYRIL BURT, 9, Elsworth Road, London, N.W.3. Authors will receive 25 copies of their articles free; extra copies may be purchased provided the order is given when the proofs are returned. Contributors should indicate whether they wish their reprints to be bound in covers. The printers will state current prices when sending out proofs.

Books for review should be sent to the Editor, advertisements to the Secretary, British Psychological Society.

The Society also issues quarterly *The British Journal of Psychology* and *The British Journal of Medical Psychology*: subscriptions, 60s. per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society publishes, jointly with the Association of Teachers in Colleges and Departments of Education, *The British Journal of Educational Psychology*: for this the subscription, £1 per annum, should be sent to Methuen & Co., Ltd., 36, Essex Street, London, W.C.2. Members of the Society receive the foregoing journals on special terms; enquiries should be addressed to the Secretary, British Psychological Society.

Papers for publication in *The British Journal of Psychology* should be sent to Professor James Drever, Department of Psychology, the University, Edinburgh; those for publication in *The British Journal of Medical Psychology* to Dr. J. D. Sutherland, 2, Beaumont Street, London, W.1.; those for publication in *The British Journal of Educational Psychology* to Professor P. E. Vernon, London University Institute of Education, Malet Street, London, W.C.1.

PREPARATION OF PAPERS

Contributors are asked to send their papers in a form which is ready for submission to the printer: that is to say, the arrangement of the material should follow that observed in previous publications of this Journal. Each article should be headed either with a series of numbered headings in italic, corresponding with the cross-headings of the several sections, or with a brief abstract: (the latter procedure is preferred; an example will be found on p. 29 of this issue). Each section should have a short cross-heading, in capitals; and the whole should conclude with a summary embodying the main conclusions reached. Special attention should be paid to such details as correct spelling, grammar, capitalization, numbering, etc. (particularly in the case of tables and references). In preparing manuscripts and in correcting proofs authors should conform, so far as possible, with the 'Recommendations to Authors' set out in *The Printing of Mathematics* by T. W. Chaundy, P. R. Barrett, and Charles Batey (Oxford University Press, 1954, ch. II, pp. 21-73).

Under the new arrangements for printing, the cost for setting up tables and algebraic formulae which require hand composition is nearly four times as much as the cost of machine composition. Contributors are therefore advised to arrange their material so that it can be printed as economically as possible. Often numerical results and algebraic formulae can, with a little simplification, be readily adapted for purposes of machine composition. Manuscripts involving many tables and displayed formulae have to be submitted first to the printers for an estimation of the cost of the various portions, and may eventually have to be returned, after some delay, to the authors for drastic revision.

Minor contributions (e.g., summaries of theses and the like) should be arranged in the form adopted for printing 'Notes'.

In future authors will receive two complete slip-proofs of their papers, but no page proof. They will be expected to pay the cost of all corrections for which they themselves are responsible.

NOTE TO SUBSCRIBERS

The Publishers wish to express their regret for the delay in the publication of the current issue: this has been due to unforeseen difficulties.



Digitized by the Internet Archive
in 2026



[By courtesy of the Psychometric Society]

PROFESSOR L. L. THURSTONE

PROFESSOR L. L. THURSTONE

British psychologists will have been deeply grieved to hear of the death, at a comparatively early age, of Professor Thurstone. He had sent a paper to the *Colloquium on Factor Analysis* held at Paris last July, but was prevented from attending personally owing to illness, and died only a couple of months later. For long the best-known leader in the field of statistical psychology, he will be mourned in every country where psychology is taught.

His parents were both born in Sweden. His father was at one time mathematical instructor in the Swedish army, but later became a Lutheran minister and an editor of a daily paper. He changed the family name, which was originally Thunström, on coming to the United States. Their only son was born at Chicago in 1887; and, owing to the movements of the family, was educated partly at Stockholm and partly in Illinois, Mississippi, and Jamestown, New York. While still at high school he won a prize of 30 dollars in the Prendergast competition in geometry; and a year later published his first printed contribution—a letter entitled ‘How to Save Niagara’, which described a way of securing 3,500,000 horsepower from the falls, without impairing their beauty.

Like Godfrey Thomson, he was originally interested in electrical engineering; and, like Boring, who was a few years his senior, he took his engineering degree at Cornell. Among his other courses he attended lectures on psychology by Titchener, whom he found (so he tells us) ‘pompous and formal’, although the subject-matter aroused his interest. Meanwhile, his own originality burst out afresh in the invention of a new type of motion-picture projector, with a film that moved continuously instead of intermittently, and so eliminated flicker. He demonstrated his model to Thomas Edison, and as a result was offered a post at Edison’s laboratory in New Jersey. Not long afterwards, however, he decided to return to academic work, and accepted an appointment in the Engineering College of Minnesota University. Here one of his pupils was Karl Holzinger.

As a result of his own experiences as a student, Thurstone for some time had been interested in the problem of learning. Accordingly, in 1914 he moved to Chicago in order to study educational psychology under Judd and general psychology under Angell. He relates that one of his own special accomplishments in learning during this period was the feat of ‘carrying five soup-plates over my head through the swinging doors at the University Commons where I worked as a waiter’. His doctor’s dissertation dealt with the curve of learning and its mathematical treatment; and not long afterwards he published a paper which contained (in the words of a recent authority¹) ‘one of the earliest important formulations of a “rational” equation’.

In 1915 Walter Bingham invited him to accept a staff-appointment at the newly established Division of Applied Psychology at the Carnegie Institute of Technology in Pittsburgh—the first department of its kind in the country. According to Bingham, Thurstone ‘took learning as his province, and initiated experiments in the printing department in co-operation with the linotype instructors’. When the United States joined in the first World War, practically the entire staff of the Institute became engaged on personnel work under the Adjutant General or the Surgeon General. Thurstone’s special assignment was to design objective methods for evaluating various oral trade tests in the Trade-Test Division at Newark, New Jersey. The success of the Army Alpha Test (largely the result of previous work at Carnegie), the popularity of the ‘National Intelligence Test’, and the use of intelligence tests in Britain, started heated discussions both here and in America about the nature of the processes so measured. Accordingly, at the first International Congress of Psychology that met after the war (Oxford, 1923) Myers organized a symposium on ‘The Nature of Intelligence’. Thurstone contributed a short paper in which he developed the thesis that ‘the degree of intelligence is associated with the degree of incompleteness of the

¹ Carl Hoyland, *ap. S. S. Stevens, Handbook of Experimental Psychology*, p. 678.

act at which it becomes focal in consciousness'. The paper formed the basis of his first book—*The Nature of Intelligence*, which was accepted by a London publisher and appeared in the International Library of Psychological Philosophy and Scientific Method (1924).

It was during the Oxford Congress, in the tiny psychological laboratory where Brown, Flügel, and I had carried out our earliest experiments on mental testing under McDougall, that my colleagues and I first met him. I well remember how impressed we were by his keen and critical mind and by the original and independent way in which he handled every problem that cropped up. In the following year we met once again at the Toronto meeting of the British Association. At its close we spent a whole day together. Thurstone joined a small party of English psychologists (including Flügel and myself) who had planned a trip to Niagara Falls. We expected to hear something of his scheme for 'saving the falls'. Instead, as we sat together on the deck of the small steamer, he began to outline his own plans for a novel approach to mental testing, which he proposed to take up at Chicago, where he had just been appointed professor. In America, he said, such courses had hitherto been virtually confined to the Binet Scale, with its clumsy rule-of-thumb standardization; in Britain Spearman and his followers seemed to be equally engrossed in statistical calculations divorced from any adequate experimental basis. His own view was that the most profitable line of attack would consist in applying the rigorous techniques already developed in the experimental laboratory, namely, the so-called 'psychophysical methods'. I for one never forgot his illuminating arguments in favour of his new proposals; and I have often wondered whether our own arguments in favour of a multi-factorial approach along Pearsonian lines then first started germinating in his fertile mind.

Between 1924 and 1928 he published a series of articles on the scaling of mental and educational tests, the mental age concept, and the growth curve for the Binet tests. They were at once recognized as the most suggestive contributions that had so far appeared. As he points out, in America the majority of those who were using tests of intelligence still seemed tacitly to assume that the distribution of scores was the same with children as it was with adults, and that 'the only effect of an increase of age was an increase in the means'. In this country, results obtained from early surveys had shown that the standard deviation tended to increase very markedly with increasing age, and that, during school years, the change was roughly linear, i.e., that the ratio of the deviation to age was approximately constant—thus furnishing some empirical support to the concept of the I.Q., or 'mental ratio', as British workers called it. As a result it proved possible to allocate the various Binet tests to points on a uniform scale by relating means to standard deviations, and so to demonstrate the great irregularities in the original sequence of tests. The conclusions drawn aroused a good deal of controversy at the time. Accordingly, with these and similar results in mind, Thurstone decided that it was necessary 'to assume *two* parameters for every age-group, namely, the mean and a measure of dispersion'. He then proposed to 'locate a rational origin for the scale' by extrapolating to the theoretical point at which the variability would vanish; and was thus led to construct his well-known 'absolute scale of intelligence'¹—a concept which, as he himself points out, was analogous to Kelvin's notion of an absolute scale of temperature. The final outcome of all this work was the booklet, invaluable to every teacher of psychology, on *The Reliability and Validity of Mental Tests* (1931): it not only presents an admirably lucid and logical statement of the whole issue, but also furnishes a plausible solution.

Meanwhile, in keeping with his original principles, he continued to devote close attention to psychophysics, which, as he writes, 'is psychological measurement proper'. The classical method of constant stimuli is, he contended, merely a special case of the more complete method of paired comparisons; and on this basis he proceeded to formulate, in precise mathematical terms, his 'law of comparative judgment'. The paper giving this equation he himself regarded as 'the best I have produced'. It was followed by half-a-dozen others proving the falsity of the phi-gamma hypothesis and seeking to bring 'some kind of system or order into psychophysics'. The method of rank order was incorporated into the same theoretical structure, and shown to be consistent with the method of paired comparisons—

¹ L. L. Thurstone, 'The Absolute Zero in Intelligence Measurement', *Psychol. Rev.*, XXXV., 1928, pp. 175-195.

an important contribution which, I fancy, has been overlooked by recent British writers on the method of ranking.

In the hope, he says, of 'introducing some life and interest into psychophysics, which had long been dominated by the trivial problems of lifted weights and limen determinations', he now attempted to extend these traditional techniques to the measurement of social values. At the time, he explains, he was corresponding with Floyd Allport about the appraisal of political and social opinions; and this eventually led to the monograph on *The Measurement of Attitudes*, written in collaboration with Professor E. J. Chave. In this field his most important work, so he himself considered, was his various studies (aided by the Payne Fund) on the effect of motion pictures on the attitudes of high school pupils. He was able, for instance, to produce 'a most convincing demonstration that the popular film called *The Birth of a Nation* had done much to make children less friendly towards the negro'.

After completing several studies on these lines, he decided that 'the construction of more and more attitude scales was becoming unproductive'; the material was 'thrown into the waste-paper basket'; and he therefore discouraged further work of this kind in his laboratory. 'I wanted', he says, 'to clear the place for developing multiple factor analysis'. Here, he explains, the turning-points 'again depended on minor incidents'. Kelley had just published his *Crossroads in the Mind of Man*, pleading for the determination of 'uncorrelated unitary traits' as an aid to 'classification, education, and guidance'; and Thurstone, in describing to a couple of colleagues over lunch his own endeavours in this direction, asked whether any mathematical techniques were available to assist in an analysis of this kind. 'They both laughed, and told me I was "extracting the latent roots of a matrix"'; and, when I asked what was meant by a matrix, they suggested that I should talk with Professor Barnard'. The outcome was the 'principal axes formula' as a method of solving a problem in least squares with a conditional equation. This, of course, was the approach that had been followed thirty years before by Karl Pearson in the papers that had inspired the earlier factorial work in this country. In applying Pearson's principles to test-data, British psychologists had introduced a succession of simplifications and changes into his procedure; and the original method and even its first author had almost dropped out of sight. Though Thurstone started once again with the full traditional procedure, he, too, quickly decided that the peculiar requirements of psychological work called for special adaptations; and it would form a most instructive study to consider in detail the points of similarity and difference between the modifications to which Thurstone in America and Pearson's critics or followers in this country were independently led, and then inquire into the reasons for the differences.

In this new field Thurstone's plans were avowedly ambitious. But they seem to have been recognized from the very outset as well worth pursuing, and he had little difficulty in securing a grant to help in developing tests for the chief primary abilities, applying them on an unusually extensive scale, and constructing a matrix multiplying machine. At the start, he says, he scarcely dreamt that he would be 'giving a major effort to this problem and its applications during the next twenty years'. Nevertheless, it is probably on the results of that major effort that Professor Thurstone's international fame now chiefly rests. He has told how he went on to undertake an identification of the primary mental abilities by the method of 'simple structure', and ultimately determined, in spite of objections from Thorndike, Kelley and others, to substitute an oblique reference frame instead of the orthogonal schemes that had previously been adopted, and finally to propose the novel concept of 'second order factors'. These were all embodied in the second edition of *The Vectors of Mind*, which was now re-christened *Multiple Factor Analysis* (1947) and remains his *magnum opus*.

In 1952 he retired from the chair at Chicago, and accepted a post at the University of North Carolina as Research Professor and Director of the Psychometric Laboratory. During the last few years of his life his interest centred chiefly in the development of performance tests of personality. Like many factorists in this country, he became greatly dissatisfied with 'the crude conceptions and the semi-intuitive procedures embodied in the popular projective tests and biographic inventories', and determined to elaborate methods more scientific. His investigations had made considerable progress; and it is to be hoped

that the material already collected and the conclusions so far reached will be published in the near future.

One feature, I think, has impressed every British student of his work. All his writings are delightfully clear and methodical. Technical terms are formally defined; new concepts carefully explained; assumptions explicitly stated; and the whole exposition developed in systematic order. Naturally the various hypotheses and inferences put forward in covering so wide a field have provoked considerable controversy; but they have also stimulated a vast amount of fruitful research. As another reviewer has noted, Thurstone's 'independence of thinking showed itself in the fact that he did not read widely in psychological literature, as he himself was quite willing to admit'.¹ Yet in many ways this has proved most fortunate. When investigators, often starting out with entirely different preconceptions independently reach identical conclusions, our confidence in those conclusions is redoubled. Several of his formulae and procedures may have been anticipated by previous writers; but no one succeeded in working them into so coherent and so attractive a scheme.

Thurstone leaves behind him many former students, who were inspired by his teaching to carry on his ideals and have achieved reputations of their own in the same field of work. It was owing to his initiative and advocacy that the Psychometric Society was formed twenty years ago; and its journal, *Psychometrika*, for whose foundation he had worked so long, and to which he made so many valuable contributions, will remain as the most lasting monument of his work. The high aim which is announced on the cover of each number might well serve as the motto of his life—'the development of quantitative rationale for the solution of psychological problems'.

CYRIL BURT

¹ J. P. Guilford, *Psychometrika*, XX, 1955, p. 264.

AGREEMENT ANALYSIS: CLASSIFYING PERSONS BY PREDOMINANT PATTERNS OF RESPONSES¹

By LOUIS L. McQUITTY
University of Illinois

I. Problem. II. Developing the Method of Agreement Analysis. III. Illustrative Applications. IV. Interpreting Results. V. Versions of Agreement Analysis. VI. Summary.

I. PROBLEM

Clinical psychology has two dissimilar heritages—dynamic psychology and psychometric methods. In harmony with the latter it has stressed objectivity [17]; in sympathy with the former, it has focused on patterns of behaviour [7]. It encountered a serious problem, however, because psychometrics has tended to neglect patterns of responses in favour of linear models [3, p. 5]. Clinicians have accordingly turned to instruments such as projective tests, that assist in assessing configurations [15]. And here clinical psychology has shown a readiness to sacrifice its birthright of objectivity to its interest in patterns. However, as psychometrics was about to lose one of its most thriving offsprings, it has broadened its perspective and capabilities by demonstrating that configurations can be objectively assessed [6]. At least one important research on psychological well-being has concluded that mental hospital patients differ from other persons primarily in terms of their patterns of responses [11].

The more highly developed pattern analytic methods, such as those by Cattell [1] or Cronbach and Gleser [5], assess profile configurations rather than patterns of responses to individual items. In the profile approaches responses to individual items are used to yield total scores on several variables; and the configurations are isolated in terms merely of patterns of standings on scales, i.e., on linear continua. Thus, they are methods for studying data ordered to linear continua; and data that do not fit are discarded.

Zubin [18] has pointed out that much information may be lost in thus allocating data to linear continua. Meehl [13] has shown that it is theoretically possible for responses treated in combination to have a predictive value which they lack when treated separately. This paradox, as he calls it, is recognized by mathematicians. They take account of it in their definitions of *independence* by stating that 'the property B_0 is said to be completely independent of properties $B_1, B_2, \dots, B_n$ if two conditions are satisfied: (i) B_0 is independent of every property $B_1, B_2, \dots, B_n$ taken separately, and (ii) B_0 is independent of the logical product of every group of properties selected out of $B_1, B_2, \dots, B_n$.' (14, pp. 57-60).

Zubin [18] in his agreement score (defined as the number of test items on which two subjects agree in their responses) has laid a foundation on which it is possible to build a pattern analytic method for classifying subjects in terms of major patterns of responses to individual items of a test. Zubin, however, did not develop the method in any general sense. It is the object of the present paper to develop and illustrate a comprehensive procedure for classifying persons in terms of their major patterns of responses. The procedure proposed is, in fact, so general that it has many versions, several of which will be outlined after the general method has been described. It will be called agreement analysis in recognition of Zubin's contribution in developing the agreement score.

¹ Appreciation is expressed to the U.S. Public Health Service for support of research out of which this method developed.



Agreement Analysis: Classifying Reason

In agreement analysis, the responses may concatenate in any fashion whatsoever: they are not restricted to linear continua; the method does not order the data according to any preconceived model. Rather, it classifies the subjects in terms of those patterns which include the greatest possible number of responses for each. These are called predominant patterns; and the data are ordered in terms of them. Responses that do not fit these patterns can be used later to reclassify the subjects in terms of less predominant patterns if it seems worth while.

Agreement analysis was not conceived originally in its most general form. A restricted version was first developed. This has been applied in two reported studies [10, 12]. Findings from these, together with the configurational theories and evidence already mentioned in support of patterns, emphasize the need for a general analytic method of classifying subjects in terms of their predominant patterns of responses, so that the reported theories and evidence may be more critically investigated.

Before developing the method, it will be helpful to outline its relation to Lazarsfeld's latent structure analysis [16, pp. 362-472]. Although this yields discrete latent classes indicated by patterns of responses, its basic psychological theory requires an underlying continuum, as Coombs [4] has pointed out. In it any given set of data may indicate one or more underlying linear continua. A person's standing on a single continuum, or his configuration of standings on more than one continuum, determines his pattern of responses. The analysis yields first the latent classes presumed to derive either from a concentration of standings on a continuum or from configurations of such standings. Secondly, it assigns to each pattern the probability with which it indicates that a subject holds membership in any latent class.

Agreement analysis differs from this procedure in not requiring a theory of underlying or latent continua. It is based on a more general postulate, namely, that there are various kinds of underlying psychological structures or predispositions (not just dimensional ones), which result in patterns of responses. Each pattern isolated by the method will be peculiar to a single class or category of subjects.

II. DEVELOPING THE METHOD OF AGREEMENT ANALYSIS

Definitions. The requisite definitions will first be stated. A *pattern of responses* consists of the responses given by one or more individuals to specified stimuli, e.g., those which he gives to the items of a psychological test. When speaking of the responses of but one individual, we call them an *individual pattern*. The analysis first classifies the individual patterns into categories called *species*. The responses common to all individual patterns of a species constitute a *species pattern*. After each individual pattern has been so classified, the species patterns are classified into broader categories of them called *genera*. The responses common to all patterns in a genus constitutes the *genus pattern*. The genera are then classified into *families*; these in turn into *orders*, *classes*, etc.: and the respective patterns of each are determined.

Assumptions. The generalized method of analysis will be developed from the following assumptions.

1. The individual patterns of responses to the items of a test are indicative of m categories of them, where m is determined by analysing the data.
2. Each response to each item is either relevant or it is irrelevant to each category. A response is relevant if, and only if, each pattern of the category includes it; otherwise it is irrelevant. If an item has a response which is relevant to a category, the item itself is also relevant to the category. If all the responses are irrelevant, then the item is irrelevant. The category pattern is the responses to the relevant items contained in the patterns classified in the category; all of these patterns have the same response for each relevant item.
3. Responses to irrelevant items are due to chance, i.e., they result from influences other than those yielding the categories into which the patterns are classified; and these other influences are assumed to be random with respect to the patterns of classification.

The Fundamental Equation. The symbols to be used may be defined as follows:

N = the number of persons in the sample to which the test has been administered.

k = the number of response alternatives for each item of the test: (all are assumed to have the same number of response alternatives).

i = any pattern at any level of classification (individual, species, genera, etc.).

j = any other pattern at any level.

$\dot{a}$ = any individual pattern which is either identical with a particular pattern i or classified in the category which yields pattern i .

p = the number of individual patterns a classified in the category which yields pattern i .

b = any individual pattern which is either identical with a particular pattern j or classified in the category which yields pattern j .

q = the number of patterns b classified in the category which yields pattern j .

r = the total number of items in the test.

n_{ij} = the agreement score for patterns i and j , i.e., the number of items for which pattern i has the same response as pattern j . (This is the agreement score developed by Zubin [18]. As pointed out by C. F. Wrigley, it is a special case of the concomitance index used by McQuitty [8], in which the denominator can be ignored because both terms remain equal throughout the study.)

E_{ij} = the total number of items which are irrelevant with respect to either the category which corresponds to pattern i or the one which corresponds to pattern j , or both (see assumptions 2 and 3, above).

n'_{ij} = the corrected agreement score for patterns i and j , i.e., n_{ij} less the number of irrelevant items on which the two patterns agree by chance.

From these definitions and assumptions, it follows that

$$E_{ij} = \frac{k^{p+q-1}}{k^{p+q-1}-1}(r-n_{ij}). \quad (1)$$

Hence

$$n'_{ij} = n_{ij} - \left(\frac{1}{k}\right)^{p+q-1} E_{ij} \quad (2)$$

$$= n_{ij} - \left(\frac{1}{k}\right)^{p+q-1} \left[\frac{k^{p+q-1}}{k^{p+q-1}-1} (r-n_{ij}) \right] \quad (3)$$

$$= n_{ij} - \frac{r-n_{ij}}{k^{p+q-1}-1}. \quad (4)$$

Eqn. (4) is fundamental for agreement analysis. In using it, we assume that the best initial indication of a species is the two individual patterns which have the highest corrected agreement score. In the above application of eqn. (4), i and j would each represent an individual pattern rather than patterns of higher level categories. Consequently, p and q would each equal one. Hence for application to two such patterns, i and j , the equation can be written:

$$n'_{ij} = n_{ij} - \frac{r-n_{ij}}{k-1}. \quad (5)$$

In the initial application of eqn. (4) (represented by eqn. (5), where the corrected scores between individual patterns are compared with those between other individual patterns) the uncorrected agreement scores would give the same classification as the corrected. This is not necessarily true later where scores between patterns are for categories classifying numbers of individual patterns. Here the amount of the correction depends in part on the number of individual patterns classified. Consequently, the corrected scores are needed, and must therefore be computed at the outset.

Agreement Analysis: Classifying Reason

Steps in the analysis. To determine which two individual patterns have the highest corrected agreement score, a matrix is computed which shows the corrected score for the agreement of each individual pattern with every other. The first tentative species will consist of individual patterns with the highest corrected agreement score. We designate this tentative species S_1 , and the individual patterns which compose it 'patterns x and y of S_1 .' After the first tentative species has been determined, its pattern (defined as in assumption 2) is written down. It consists of the responses to the items for which any two individual patterns agree. This species pattern is designated SP_1 .

Using eqn. (4), the corrected agreement score for each individual pattern with SP_1 is computed. Let j represent the species pattern and i an individual pattern; then $q = 2$ and $p = 1$; and the equation can be written

$$n'_{ij} = n_{ij} - \frac{r - n_{ij}}{k^2 - 1}, \quad (6)$$

where n_{ij} — the number of items on which the species pattern j has the same response as the individual pattern i .

A special case arises when the corrected agreement score is desired between pattern j and either individual pattern x or y . Suppose that the score desired is between pattern j and individual pattern x , denoted n'_{xj} . An estimate denoting $\hat{n}_{xj}$ could be obtained from the analogous uncorrected score n_{xj} , if we had an estimate of the latter. An estimate of the latter $\hat{n}_{xj}$, will be n_{xy} , so that $\hat{n}_{xj} = n_{xy}$.

From this by analogy with eqn. (4), we have:

$$\hat{n}_{xj} = n_{xy} - \frac{r - n_{xy}}{k^{p+q-1} - 1} \quad (7)$$

Eqn. (7) is used to compute the corrected agreement score between an individual pattern x and the category in which it is classified, provided that this contains only one other pattern y . In the general case where two patterns i and j' have joined to form a category which yields pattern j , and where i is so assigned that its category contains either fewer or the same number of individual patterns as the category corresponding to pattern j' , the following equation applies by analogy:

$$\hat{n}_{ij} = n_{ij'} - \frac{(r - n_{ij'})}{k^{p+q-1} - 1} \quad (8)$$

As eqn. (6) and (7) are applied to determine the corrected agreement score of each individual pattern with SP_1 , the original matrix of agreement scores between individual patterns will be expanded to include them. When all these corrected agreement scores with SP_1 have been added to the matrix, the largest score mediating between either (a) two unclassified individual patterns or (b) SP_1 and an unclassified individual pattern will be selected. If it is between two unclassified individual patterns, a new tentative species has been isolated, S_2 , leading to a new tentative species pattern, SP_2 . If, however, it is between SP_1 and an unclassified individual pattern, then S_1 has been expanded and SP_1 more definitely determined.

For the purpose of outlining the method further, let us suppose that the largest corrected agreement score is between two as yet unclassified individual patterns to give S_2 and SP_2 , with these two newly classified patterns now designated as individual patterns x and y of S_2 . Using eqn. (6) and (7), we determine the corrected agreement score for each individual pattern with SP_2 , and add these to the matrix of agreement scores. We also compute the corrected agreement score between SP_1 and SP_2 . Eqn. (4) may be adapted for this purpose as follows:

$$n'_{ij} = n_{ij} - \frac{r - n_{ij}}{k^{2+2-1} - 1} \quad (9)$$

In eqn. (9), either i or j represents SP_1 and the other SP_2 . The corrected agreement score for these two is then added to the matrix, and the largest corrected agreement score is chosen from those in the following categories: (a) between unclassified individual patterns, (b)

between an unclassified individual pattern and a species pattern and (c) between the two species patterns.

We may further suppose that the largest score is between an individual pattern and SP_2 , giving an additional individual pattern, z in S_2 . As thus enlarged, S_2 is designated S_2^1 with a species pattern SP_2^1 . The definition of the species pattern has now been changed by the addition of another individual pattern. It may have changed so much that individual pattern x or y of S_2 may no longer classify best with a pattern of the species. To determine whether or not this is so we apply a criterion of internal consistency of classification. In doing this we determine the corrected agreement score of SP_2^1 with individual patterns x and y using eqn. (8). We then search all of the corrected agreement scores in which x and y participate, to discover if any of the patterns has a larger score with a pattern not in S_2^1 than it does with the pattern of S_2^1 . If so, an inconsistency has been found. To solve this inconsistency, individual pattern z is kept out of S_2 (and its derivatives) until it can enter without causing an inconsistency. If this never occurs, the pattern must be treated as a species containing only a single case.

After all the individual patterns have been classified into species the method of analysis is repeated at the next higher level; species patterns are classified into genera in the same manner as individual patterns were classified into species. Here, adaptations of eqn. (4), analogous to the adaptation in eqn. (9) will be used. Eqn. (9) represents an adaptation for computing the corrected agreement score between two species where each species classifies two individual patterns, and consequently both the p and q of eqn. (4) are given as 2 in eqn. (9). Otherwise, they will be listed as the number of patterns classified under the categories corresponding to the two patterns, i and j respectively. After the classification of species into genera has been completed, then genera are classified into families, and so on, until at the end either there are only two top level categories, and each next lower level pattern is classified into one of them, or the classification proves to be complete because all corrected agreement scores above zero have been exhausted.

III. ILLUSTRATIVE APPLICATION

To clarify the method just described, it will now be applied to the responses of 9 subjects to the 36 collective items of a personality inventory [9]: e.g., (i) Do you often borrow money? (ii) Are you very talkative? In answering them Y is written for 'yes', N for 'no' B for 'between the two', i.e., doubtful.

The 9 subjects were selected from a group of 80 adult Negro males, all normal. In selecting the nine, each of the 80 adults was first paired with every other. The agreement score was determined for each pair. The five pairs with the largest scores were selected. However, to make the problem of classification harder, one of the individuals with the fifth largest score was omitted; this left only 9. Their responses are shown in the first 9 columns of Table I. The responses of each person constitute an individual pattern, labelled a, b, c, d, e, f, g, h , or i . The steps in the analysis are as follows:

1. Each individual pattern is paired with every other. For each pair we first count the number of items for which the two patterns agree in the answers. For example, a and b disagree on 9 answers (1, 6, 8, 9, 13, 20, 22, 23, and 29). They agree in their answers to all the rest. Their agreement score is therefore $36 - 9 = 27$. These scores are entered in the first 9 rows and columns of Table II. Diagonal entries are not required.

2. Using eqn. (5), the entries in Table II are converted to corrected agreement scores. Now, there are, it will be remembered, 3 alternative responses; however, the third ('B') was only used in about 11 per cent of the answers, so that the test functioned virtually as a 2-response test: and we may therefore give k a value of 2 throughout this example. Accordingly, taking the pair a and b and substituting the numerical values shown in the table, we obtain $n'_{ij} = 27 - \frac{36-27}{2-1} = 18$. This result is entered in the appropriate cell of Table III. Since h omits responses to 2 of the items, the test was treated as containing only 34 wherever pattern h was involved.

3. The highest n'_{ij} (22 for patterns a and h) is selected. We assume that this value represents the best evidence for a tentative species, S_1 , composed of the two individual patterns a and h .

Agreement Analysis: Classifying Reason

TABLE I. INDIVIDUAL AND CATEGORY PATTERNS OF RESPONSES TO 36 ITEMS

Items											Patterns							
	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>i</i>		(<i>ah</i>) <i>SP</i> ₁	(<i>ahg</i>) <i>SP</i> ₁ ¹	(<i>cf</i>) <i>SP</i> ₁	(<i>ahgb</i>) <i>SP</i> ₁ ²	(<i>di</i>) <i>SP</i> ₃	(<i>ahgbe</i>) <i>SP</i> ₁ ³	<i>ghb</i>	<i>GP</i> ₁ ¹
1	N	Y	N	N	N	Y	N	N	N		N	N			N			
2	N	N	N	N	N	Y	N	N	N		N	N	N	N	N	N	N	N
3	N	N	N	N	N	B	Y	N	N	*	N							
4	Y	Y	Y	B	Y	Y	Y	Y	Y	B	Y	Y	Y	Y	B	Y	Y	Y
5	Y	Y	N	N	N	Y	N	N	N	Y	Y							
6	N	Y	N	N	N	Y	N	N	N	Y	Y	N						
7	N	N	N	Y	N	N	N	N	Y	Y		N			Y			
8	Y	N	Y	Y	B	Y	N	N	N	Y	Y		Y	Y	Y			
9	N	Y	N	B	B	Y	N	N	N	B	N	N			B			
10	Y	Y	Y	Y	Y	N	N	Y	Y	B	Y	Y					Y	
11	Y	Y	N	Y	Y	N	N	Y	Y	Y	Y	Y		Y		Y	Y	
12	N	N	N	N	N	N	N	N	Y	N			N		Y			
13	N	B	Y	N	B	N	N	N	N	B	N		N					
14	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
15	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
16	N	N	N	N	B	Y	Y	Y	Y	Y	Y							
17	Y	Y	B	B	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	
18	Y	Y	N	B	B	N	N	Y	Y	Y	Y	Y	N	Y		Y	Y	
19	N	N	N	B	N	N	N	N	Y	Y	N	N		N		N	N	N
20	Y	N	Y	Y	B	Y	Y	Y	Y	Y	Y		Y		Y			
21	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
22	Y	N	N	B	N	Y	Y	Y	Y	B	Y				B			
23	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N			
24	N	N	N	N	B	Y	N	N	N	N	N			N			N	
25	Y	Y	B	N	N	Y	N	N	N	Y								
26	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y		Y		Y			
27	N	N	N	N	Y	N	N	N	N	B	N	N	N	N			N	
28	N	N	N	B	N	N	N	N	N	N		N			N			
29	Y	N	N	Y	B	N	B	N	N	Y			N		Y			
30	N	N	N	N	B	Y	N	N	N	B	N	N		N			N	
31	Y	Y	Y	N	Y	N	N	B	N	N								
32	N	N	Y	N	N	Y	N	N	N	B	N	N	Y	N		N	N	
33	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
34	N	N	B	N	B	N	Y	N	N	N	Y				N			
35	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
36	Y	Y	Y								Y				Y			

* No answer given by the subject.

TABLE II. UNCORRECTED AGREEMENT SCORES BETWEEN INDIVIDUAL AND CATEGORY PATTERNS

Persons											(<i>ah</i>) <i>SP</i> ₁	(<i>ahg</i>) <i>SP</i> ₁ ¹	(<i>cf</i>) <i>SP</i> ₂	(<i>ahgb</i>) <i>SP</i> ₁ ²	(<i>di</i>) <i>SP</i> ₃	(<i>ahgbe</i>) <i>SP</i> ₁ ³	(<i>ahgbecf</i>) <i>GP</i> ₁ ¹
	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>i</i>								
<i>a</i>																	
<i>b</i>	27		27	25	23	20	25	27	28	23	28	23	18	16	16	11	
<i>c</i>	25	21		21	17	17	21	21	21	20	20	16	15	16	12	11	
<i>d</i>	25	21	20		20	19	22	22	21	15	20	17	22	12	13	8	
<i>e</i>	23	17	20	17		17	18	23	22	21	19	14	14	11	21	8	
<i>f</i>	20	17	19	17	15		15	20	18	15	17	15	12	11	12	11	8
<i>g</i>	25	21	22	18	15	20		22	19		20	16	22	11	14	9	
<i>h</i>	27	21	22	23	20	20	25		18	23	23	15	16	16	14	11	
<i>i</i>	28	21	21	22	18	22	25	21		28	23	16	16	15	15	11	
	23	20	15	21	15	19	18	21		18	14	12	11	21	8		
(<i>ah</i>) <i>SP</i> ₁	28	20	20	19	17	20	23	28	18			23	15	16	14	11	
(<i>ahg</i>) <i>SP</i> ₁ ¹	23	16	17	14	15	16	23	23	14	23			13	16	11	11	
(<i>cf</i>) <i>SP</i> ₂	18	15	22	14	12	22	15	16	12	15	13			10	10	8	8
(<i>ahgb</i>) <i>SP</i> ₁ ²	16	16	12	11	11	11	16	16	11	16	16	10			8	11	8
(<i>di</i>) <i>SP</i> ₃	16	12	13	21	12	14	14	15	21	14	11	10	8			7	6
<i>SP</i> ₁ ³ = <i>GP</i> ₁ ¹ (<i>ahgbe</i>)	11	11	8	8	11	9	11	11	8	11	11	8	11	7			8
<i>GP</i> ₁ ¹ (<i>ahgbecf</i>)					8								8	8	6		

4. SP_1 , the pattern for this species, is shown in the tenth column of Table I. It consists of the responses for the items on which patterns a and h agree.

5. Using eqn. (6), the corrected agreement score for SP_1 with each individual pattern not in S_1 is now computed. For this purpose we need the agreement score for SP_1 with each of these individual patterns. These are obtained from Table I by counting the number of agreements between SP_1 and each individual pattern, and are set out in the SP_1 column and row of Table II. Thus to calculate the corrected agreement between SP_1 and pattern b we find $n_{ij} = 30$, and $r = 34$ (because the person represented by pattern h of S_1 answered only 34 items). Thus $n'_{ij} = 20 - \frac{34-20}{2^2-1} = 15.33$. The corrected agreement scores so obtained are entered in the SP_1 row and column of Table III.

6. Eqn. (7) is used to compute the corrected agreement score between SP_1 and each individual pattern a and h classified in S_1 . Here x in the equation represents a and y represents h ; $r = 34$ (as before) and $q = 2$ because S_1 contains two individual patterns. Thus

$$n'_{ij} = n_{ab} - \frac{34 - n_{ah}}{2^2 - 1} = 28 - \frac{34 - 28}{3} = 26.$$

This and similar results are entered in the appropriate cells of the SP_1 row and column of Table III.

TABLE III. CORRECTED AGREEMENT SCORES BETWEEN INDIVIDUAL AND CATEGORY PATTERNS

Persons	a	b	c	d	e	f	g	h	i	(ah) SP_1	(ahg) SP_1^1	(cf) SP_3	$(ahgb)$ SP_1^1	(di) SP_2	GP_1 $(ahgbe)$ SP_1^1 GP_1^1	
a		18	14	10	4	14	18	22	10	26.00	21.43	12.00	14.80	9.33	10.26	
b	18		6	-2	-2	6	6	8	4	15.33	13.43	8.00	14.80	4.00	10.26	
c	14	6		4	2	8	8	8	-6	15.33	14.57	17.33	10.53	6.33	7.16	
d	10	-2	4		-2	0	10	10	6	14.00	11.14	6.67	9.33	16.00	7.16	
e	4	-2	2	-2		-6	4	2	-6	11.33	12.29	4.00	9.33	4.00	10.26	7.79
f	14	6	8	0	-6		4	10	2	15.33	13.43	17.33	9.33	6.67	8.81	
g	18	6	8	10	4	4		16	0	19.67	21.43	8.00	14.80	6.67	10.26	
h	22	8	8	10	2	10	16		8	26.00	21.43	10.00	14.80	8.67	10.26	
i	10	4	-6	6	-6	2	0	8		12.67	11.14	4.00	9.33	16.00	7.16	
$(ah)SP_1$	26.00	15.33	15.33	14.00	11.33	15.33	19.67	26.00	12.67		21.43	12.29	14.80	11.14	10.26	
$(ahg)SP_1^1$	21.43	13.43	14.57	11.14	12.29	13.43	21.43	21.43	11.14	21.43		11.60	14.80	9.47	10.26	
$(cf)SP_2$	12.00	8.00	17.33	6.67	4.00	17.33	8.00	10.00	4.00	12.29	11.60		9.23	6.29	7.57	7.79
$(ahgb)SP_1^1$	14.80	14.80	10.53	9.33	9.33	9.33	14.80	14.80	9.33	14.80	14.80	9.23		7.16	10.26	7.79
$(di)SP_2$	9.33	4.00	6.33	16.00	4.00	6.67	6.67	8.67	16.00	11.14	9.47	6.29	7.16		6.56	5.95
$GP-(ahgbe)SP_1^1$	10.26	10.26	7.16	7.16	10.26	8.81	10.26	10.26	7.16	10.26	10.26	7.57	10.26	6.56		7.79
$(ahgbecf)GP_1$						7.79										

7. In Table III we now select the largest corrected agreement score which mediates between either (a) two unclassified individual patterns or (b) SP_1 and an unclassified pattern, viz., 19.67 (between SP_1 and g). This suggests that pattern g should be classified in S_1 . The latter species is then redesignated S_1^1 and has a species pattern designated S_1^1 (the superscripts indicating that S_1 has been revised once).

So far we have determined the correct agreement score that various patterns have with individual pattern g when it is not classified with them. More critical indices are those it has with the resultant species pattern if it is classified with them: evidently, if g is to be classified in S_1 , it must have a larger corrected agreement score with SP_1^1 than with any other species pattern that might result from classifying individual pattern g with some individual pattern not in S_1 . Of all the individual patterns not in S_1 , g has its highest corrected agreement score with d , viz., 10.00, as shown in Table III. Hence, were it classified with d , it would have a higher corrected agreement score with the resultant species pattern than with any that might be obtained on classifying g with any other individual pattern not in S_1 .

Agreement Analysis: Classifying Reason

Using Eqn. (7), we find that g would have a corrected agreement score of 18.67 with the tentative species pattern resulting from classifying individual patterns d and g together. This score is smaller than g has with SP_1^1 . It is therefore smaller than g would have with SP_1^1 , as can be seen by comparing eqn. (4) and (8). Consequently, individual pattern g is classified in S_1^1 to produce S_1^1 and SP_1^1 . The pattern SP_1^1 is shown in Table I; it consists of the responses common to SP_1 and individual pattern g .

8. The corrected agreement scores between SP_1^1 and all other patterns are next determined. For this purpose, it is necessary to count the number of agreements between SP_1^1 and the individual patterns, as listed in Table I. These scores are shown in the SP_1^1 row and column of Table II. But to explain the further use of eqn. (4), let us apply it to compute the corrected agreement score between SP_1^1 and b . Here i and j will represent b and SP_1^1 respectively; then $p = 1$, and $q = 3$, for i represents a single individual pattern, and j represents one which is derived from 3 individual patterns. $n_{ij} = 16$ (the agreement score between SP_1^1 and b , as shown in Table II; $r = 34$ (for h answered only 34 items) and $k = 2$). On substituting these values we obtain $n'_{ij} = 16 - \frac{34-16}{2^{1+3-1}-1} = 12.43$. This and similar values for patterns not classified in S_1^1 have also been entered in Table III.

To estimate the corrected agreement score between SP_1^1 and each of the individual patterns in S_1^1 , we use eqn. (8). By way of illustration let us compute the $\hat{n}_{ij}$ between SP_1^1 and a . Here i represents the individual pattern a ; j' represents the pattern of responses common to individual patterns h and g ; $n_{ij'}$ is the number of agreements between individual pattern a and pattern j' ; this is the same as the number of agreements that individual pattern a has with SP_1^1 , viz., 23 (as shown in Table II); $r = 34$; $a = 3$ (since pattern j is for a species containing three individual patterns); $p = 1$ (since x is for an individual pattern). Substituting these values in eqn. (8), we obtain $\hat{n}_{ij} = 23 - \frac{34-23}{2^3-1} = 21.43$.

The corrected agreement scores for SP_1^1 with other patterns, individual and species, in S_1^1 are necessarily equal to the value thus obtained, as can be seen from eqn. (8). Accordingly this value is entered in the appropriate cells of the SP_1^1 row and column of Table III.

9. The next step is to search for inconsistencies of classification and make the necessary re-adjustments (if any). Such a search is needed at this stage because SP_1 , on being revised to SP_1^1 , may have changed so much that a or h may now have a higher corrected score with some pattern not in S_1^1 than with SP_1^1 . On examining rows a and h in Table III, we find that both of these patterns have higher scores with SP_1^1 than with any pattern not in SP_1^1 . Thus no inconsistency is indicated.

10. In Table III the largest corrected agreement score which mediates between either (a) two unclassified individual patterns or (b) between SP_1^1 and an unclassified individual pattern was now selected, viz., 14.57 (between SP_1^1 and c). This indicates that individual pattern c is the next to be classified. We must next determine whether it would classify better in the species that would result from combining it with SP_1^1 or in the species that would result from combining it with some individual pattern not in S_1^1 . Of patterns not in S_1^1 , c has its highest corrected agreement score, viz., 8 (Table III), with individual pattern f (Table III). Hence, if individual patterns c and f were combined to form a species, S_2 , with species pattern SP_2 , pattern c would have a larger corrected agreement score with SP_2 than with any other species pattern formed by combining it with some individual pattern not in S_1^1 . Using eqn. (8), individual pattern g 's corrected agreement score between g and SP_2 proves to be 17.33. This is a larger corrected agreement score than individual pattern c could have with the species pattern resulting from combining it with SP_1^1 , since c has an uncorrected agreement score of only 17.00 with SP_1^1 . Hence, c is classified with f to yield a tentative species S_2 .

11. A search is now made for evidence of inconsistencies of classification. None is found.

12. In Table III the largest corrected agreement score is next selected from those that mediate between the following patterns: (a) two unclassified individual patterns; (b) either SP_1^1 or SP_2 and an unclassified individual pattern; (c) SP_1^1 and SP_2 . It proves to be the score between SP_1^1 and b , viz., 13.43. This indicates that b is the next to be classified. Were it classified in S_1^1 it would have a corrected score of 14.80 with the resultant species pattern (as shown by eqn. (8)); were it classified with i (the unclassified individual pattern with which it has its highest corrected agreement score)¹ it would have a corrected score of 14.67 (as shown by eqn. (7)). Accordingly b is classified in S_1^1 to yield S_1^2 and SP_2^2 .

13. A search is again made for evidence of inconsistencies of classification. None is found.

14. The corrected agreement scores are accordingly computed for SP_2^2 , with every other pattern and entered in Table III.

15. The largest corrected agreement score is next selected from those mediating between the following patterns: (a) unclassified individual patterns; (b) SP_2 or SP_1^2 and an unclassified individual pattern; (c) SP^2 and SP_1^2 . This is 9.33 (between SP_1^2 and the patterns d, e , and i). It is now necessary to determine for each of these whether they would have a larger corrected agreement score with the species pattern resulting from classifying them in S_1^2 than with the one which results if they were classified with some unclassified pattern. The largest corrected agreement scores prove to be 6.00 (between pattern d and i). When these are grouped together to form a species, each has an agreement score of 16.00 with the pattern for the resultant species. This is larger than their agreements with SP_1^2 (Table II). Hence they are classified to form a further species, S_3 , with its corresponding pattern.

16. The corrected agreement scores for SP_3 with every other pattern is now computed and entered in Table III.

17. A search is made for evidence of inconsistencies of classification. None exists.

18. We now select the largest corrected agreement score mediating between e , SP_1^2 , SP_2 , and SP_3 , viz., of 9.33 (between e and SP_1^2): e has therefore to be classified in SP_1^2 to form S_1^2 and SP_1^2 .

19. The corrected agreement scores for all patterns with SP_1^2 are computed and entered in Table I.

20. On examining Table III we now discover evidence for three inconsistencies, all pertaining to a . This pattern, classified in S_1^2 , has a higher score with each individual pattern c and f and SP_2 (derived from c and f) than with SP_1^2 . Evidently e cannot be classified in S_1^2 to produce SP_1^2 without at the same time misclassifying pattern a . We have now therefore to ascertain which of these two patterns, a and e , classifies best with the species pattern composed of individual patterns h, g and b (the remaining members of S_1^2). This species pattern is shown in Table I: e has only 11 agreements with it while a has 16. Consequently a classifies best with it; moreover e cannot be classified elsewhere, and it must therefore be regarded as a species pattern restricted to itself. The species pattern resulting from the analysis are SP_1^2 , SP^2 , SP_3 , and pattern e . These have next to be classified into genera.

21. The highest corrected agreement score between any two species is 9.33 (between e and SP_1^2). These are therefore combined to form genus G_1 , and the SP_1^2 columns of Tables I, II, and II are accordingly given the additional label of GP_1 . The classification of pattern e with SP_1^2 to produce a genus is of course permissible, though it was not permissible to redefine a species pattern.

22. The highest corrected agreement score is now selected from the following: (a) two unclassified species pattern; and (b) GP_1 and an unclassified species pattern. It proves to be 7.57, between GP_1 and SP_2 , which are accordingly grouped to yield G_1^2 and GP_1^2 .

23. The corrected agreement scores are next computed for GP_1^2 with all the species and genus patterns.

24. A search is made for inconsistencies of classification, and none is found.

The classification is now complete, and results in G_1^2 and G_2 with patterns GP_1^2 and GP_2 . GP_2 is obtained by accepting SP_3 as defining a genus pattern because there was no genus pattern other than GP_1^2 for it to classify with. At the next higher level, GP_1^2 would come together with GP_2 to form the single family F_1 .

IV. INTERPRETING RESULTS

Predominant Patterns. It can be shown that agreement analysis classifies in such a manner as to maximize the corrected agreement score that each pattern has with the category pattern of the category into which it is classified. Consequently, the categories have the largest possible number of responses in their patterns. Categories and patterns of this kind may be called 'predominant'.

Now let us define the following symbols:

i and j are to have their customary meanings, except that they are for patterns at the same level as j' (see below);

b = any pattern classified in the category which produced any category pattern j ;

j' = the category pattern for any combination of patterns (all at the same level of classification) which did not occur in the final classification but which includes pattern b ;

j'' = any one of patterns j and j' .

It follows that—

$$n'_{ai} > n'_{aj'}, \text{ otherwise an inconsistency of classification would exist;} \quad (10)$$

$$n'_{ai} > n'_{ab}, \text{ otherwise an inconsistency of classification would exist;} \quad (11)$$

$$n'_{ab} > n'_{aj'}, \text{ since pattern } j', \text{ containing pattern } b, \text{ could not have any responses not in pattern } b; \quad (12)$$

$$n'_{ai} > n'_{aj'} \text{ and } n'_{ai} > n'_{aj''}. \quad (13)$$

This last inequality means that each pattern has an agreement score with the pattern of the category in which it is classified at least as large as it could have in any other category that might arise from any combination of patterns at the same level. Consequently, the category patterns have the maximum number of characteristic responses possible, subject to the requirements (a) that each pattern be classified with at least one other at each level (unless this leads to an inconsistency), and (b) that each pattern classified in a category have all the characteristics of the category pattern. It follows that the categories are both predominant and pure.

Describing typological differentiae. The pattern of responses for each category is assumed to indicate an inferred construct, called a typological differentia, which causes the pattern of answers to occur. Hence, each typological differentia may be described in terms of its pattern of responses. In developing these descriptions, the investigator attempts to decide what unitary characteristic would be indicated by each category pattern. His description of any one differentia is not merely a combination of single characteristics, i.e., one for each response of the pattern; he has to keep in mind the possibility that many other responses might also have patterned with those used had other items been included in the test. Consequently, it might be possible to gain further insight into the nature of a psychological differentia by interviewing subjects characterized by it. These impressions obtained from such interviews could then be objectively investigated by an agreement analysis of a test that included items designed to investigate them along with marker variables chosen from those which originally isolated the typological differentiae.

V. VERSIONS OF AGREEMENT ANALYSIS

Agreement analysis may take various forms : e.g., (i) those derived from its application to various kinds of data, and (ii) shortened versions such as may be derived by omitting certain steps of the complete method or by placing certain restrictions on the typological differentiae so isolated.

(i) *Versions arising from Difference of Data.* These exhibit an obvious analogy with the various forms of factor analysis. Cattell [2] has outlined six factorial techniques, *R*, *Q*, *P*, *O*, *S*, and *T*, obtained by applying factor analysis to various kinds of data. *R*-technique, for example, is obtained by applying it to correlations between a small number of tests procured from a large number of persons; *Q*-technique from applying it to correlations between a small number of persons subjected to a large number of tests; and so on.¹ The

It should perhaps be explained that in British factorial studies what is here called ' *Q*-technique' is more frequently termed ' *P*-technique', since the phrase ' *Q*-technique' was originally used by Dr. Stephenson to designate a somewhat specialized procedure which he himself devised. According to Burt's theory of factorial methods the six techniques mentioned in the text are derived from the six 'one-way first order interactions', arising in ordinary multivariate analysis of variance, and may consequently be applied to 'qualitative data' of the type here envisaged : (see *Factors of the Mind*, 1940, pp. 169 f., and this *Journal*, VIII, pp. 110 f.).

version of agreement analysis here outlined is analogous to *Q*-technique, since the agreement score forms an index of agreement between persons, as determined from their responses to tests. For every one of the factorial techniques, *R*, *Q*, *P*, *O*, *S*, and *T*, there will be an analogous version of agreement analysis. These are secured by applying the methods here described to the classes of data to which the corresponding factorial technique is applied.

(ii) *Versions obtained by restricting the Complete Method.* Various shortened versions of agreement analysis may be recognized. They can be recommended for isolating predominant categories only if the data meet certain criteria which will ensure that the shortened

methods give similar results to the complete method. Some of the criteria have already been recognized and will be outlined here. Others may doubtless be developed, as both shortened and complete procedures are applied to the same data. One of the shortest versions is obtained when uncorrected agreement scores are used instead of corrected scores, and no search is made for inconsistencies. When the data fall into sharply distinguishable categories, the two versions yield the same results.

Criteria for determining whether or not distinct categories exist, can be specified for both binary and multiple categories. Binary categories contain only two patterns. In this case the binary principle of reciprocity may be applied: this states that pattern i has its highest agreement score with pattern j if j has its highest with i .

Multiple categories include two or more patterns. For these an analogous principle of reciprocity can be formulated: if pattern i has its p highest agreement scores with patterns 1, 2, 3, ..., p (including pattern i), then any other pattern i' of the group has its p highest agreement scores with these same p patterns.

If either of these two 'reciprocity principles' is fulfilled for the uncorrected agreement scores, then neither corrected agreement scores nor a search for inconsistencies is required in order to isolate the predominant categories. A study in the field of union-management relations found the binary principle fulfilled [10, pp. 439-474]. Even if the principles do not hold for uncorrected agreement scores, they may still hold for corrected scores, and, if so, then a search for inconsistencies is not required to obtain predominant categories. There is another justification for using a shortened method. One might wish to classify persons in relation to an external criterion. This can be done by forcing the patterns into binary types, and, according to the results of a study in progress, a high degree of correct classification in cross validation in relation to an external criterion can still be achieved.

VI. SUMMARY

Clinical psychologists have long been interested in patterns of responses, and psychometric research has recently given attention to individual differences and similarities in configuration of scores. However, the more highly developed of these quantitative methods have restricted themselves to configurations of standing on linear continua. Considerable information may be lost in allocating data to linear continua, and configurations of responses to the individual items of a test may possess predictive values that are missed when the responses are located singly or jointly to produce scores on linear continua.

This paper develops and illustrates a general method of pattern analysis, called agreement analysis, for classifying persons according to their predominant patterns of responses to the individual items of a test. The method is general in the sense that it has many versions. One of the restricted versions was developed prior to the general method and applied in two empirical studies. Results of these studies, together with theories of the significance of patterns and other evidence in support of them, demand a generalized method for researches into the significance of patterns of responses. The method of agreement analysis here developed, makes it possible to investigate, more critically than heretofore possible, whether or not responses to individual items treated from a configural point of view do in fact have predictive values not evident when the same responses are analysed in terms of linear continua exclusively, or even in terms of configurations of standing on linear continua.

REFERENCES

1. Cattell, R. B. (1949). ' r_p and other coefficients of pattern similarity.' *Psychometrika*, XIV, 279-298.
2. Cattell, R. B. (1952). 'The three basic factor analytic designs—their inter-relations and derivatives.' *Psychol. Bull.*, XLIV, 499-520.
3. Coombs, C. H. (1952). 'A theory of psychological scaling.' *Engineering Research Bulletin*, no. 34. Ann Arbor: Univ. of Michigan Press.
4. Coombs, C. H. *Theory and methods of social measurement*. In Festinger, L., and Katz, D. (1953). *Research methods in the behavioral sciences*. New York: The Dryden Press, pp. 471-535.

Agreement Analysis: Classifying Reason

5. Cronbach, L. J., and Gleser, G. C. (1953). 'Assessing similarity between profiles.' *Psychol. Bull.*, L, 456-473.
6. Gaier, E. L., and Lee, Marilyn C. (1953). 'Pattern analysis: the configural approach to predictive measurement.' *Psychol. Bull.*, L, 140-148.
7. Louttit, C. M. (1939). 'The nature of clinical psychology.' *Psychol. Bull.*, XXXVI, 361-389.
8. McQuitty, L. L. (1938). 'An approach to the measurement of individual differences in personality.' *Charact. and Pers.* VII, 81-95.
9. McQuitty, L. L. (1952). 'Effective items in the measurement of personality integrations: I.' *Educ. psychol. Measmt.*, XII, 117-125.
10. McQuitty, L. L. 'Pattern-analysis—A statistical method for the study of types.' In W. E. Chalmers, M. K. Chandler, L. L. McQuitty, R. Stagner, D. E. Wray, and M. Derber (1954). *Labor management relations in Illini City*, II. Champaign, Illinois: Institute of Labor and Industrial Relations, University of Illinois.
11. McQuitty, L. L. (1954). 'Theories and methods in some objective assessments of psychological well-being.' *Psychol. Monogr.*, LXVIII, No. 14 (whole no. 385).
12. McQuitty, L. L. (1954). 'Pattern analysis illustrated in classifying patients and normals.' *Educ. and psychol. Measmt.*, XIV, 598-604.
13. Meehl, P. E. (1950). 'Configural scoring.' *J. consult. Psychol.*, XIV, 165-171.
14. Neyman, J. (1950). *First course in probability and statistics*. New York: Holt.
15. Sargent, H. (1945). 'Projective methods: their origin, theory, and application in personality research.' *Psychol. Bull.*, XLII, 257-293.
16. Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., and Clausen, J. A. (1950). *Measurement and prediction*. Princeton, New Jersey: Princeton Univ. Press.
17. Watson, R. I. (1953). 'A brief history of clinical psychology.' *Psychol. Bull.*, L, 321-346.
18. Zubin, J. (1938). 'A technique for measuring like-mindedness.' *J. abnorm. soc. Psychol.*, XXXIII, 508-516.

AGREEMENT ANALYSIS

A NOTE ON PROFESSOR McQUITTU'S ARTICLE

By H. E. WATSON

Factor Analysis as a Classificatory Procedure. Professor McQuitty's paper on classifying persons by predominant patterns of response will be welcomed by clinical psychologists in this country because the problem with which he deals has come to the fore in many recent investigations. In an inquiry which my colleagues and I have been carrying out it was desirable to distinguish between potential delinquents and potential neurotics, and, so far as possible, to classify each group into the main subtypes. Like Professor McQuitty we quickly found that the assessments obtained with current projection tests were decidedly lacking in objectivity. On the whole, it appeared that the answers to a personality questionnaire (we used a shortened adaptation of that printed in Burt's *Subnormal Mind*) had a much greater reliability; and we therefore attempted to determine what kind of 'discriminant function', based on these replies, would provide the best diagnostic criteria.

The procedure adopted was that outlined in an earlier issue of this *Journal* ('The Factorial Analysis of Qualitative Data', III, p. 166-185). It has since been suggested to us that it might be instructive to re-analyse our own data by Professor McQuitty's method. The essential results remain much the same as before; but certain additional points have emerged with the newer method that remained undetected with the factorial procedure, and conversely the factorial procedure has brought to light other features in the data that were not elicited by Professor McQuitty's method. In particular, the factorial procedure enables us to classify the items or traits as well as the persons assessed, and so to secure a concrete specification for each of the types into which the persons are classified. Accordingly we thought it might be of interest to try the same procedure with Professor McQuitty's own data, in order to ascertain whether or not these impressions were confirmed.

Working Methods. In comparing two variables the simplest and most convenient measure of 'agreement' is a coefficient based on their product-sum (e.g., the covariance, standardized if necessary to yield a coefficient of correlation). Burt's method therefore begins by entering 'yes's' as +1's, 'no's' as -1's, and 'between' or 'doubtful' as 0. Incidentally, this seems to yield a more satisfactory method of treating the last type of answer than that adopted by Professor McQuitty: indeed, with this numerical device it becomes possible to deal with questionnaires, etc., where multiple, and not merely binary, alternatives are allowed, whereas his procedure, apparently, requires every questionnaire to 'function as a two-response test' (p. 9). For any pair of columns or rows in Table I (p. 10) the product-sum will now furnish a measure of agreement which lends itself readily to mathematical manipulation; and it will be found that, except in the instances where the candidate answers 'between', the figure obtained is precisely the same as Professor McQuitty's 'Uncorrected Agreement Score' (Table II).

In form, the equation for 'correcting' these raw scores (cf. this *Journal*, III, p. 174) is identical with Professor McQuitty's, viz., $n_{ij} - n'_{ij} - E_{ij}$, where E_{ij} denotes the value expected in the absence of any agreement. But the method of estimating E_{ij} is different. Most British writers start with the formula first proposed by Pearson for dealing with contingency tables, namely, $n_i n_j / n$. Moreover, in the paper just cited, the resulting values are standardized: so that, where Professor McQuitty uses equation (4), Burt uses the formula $r'_{ij} = \frac{1}{\sqrt{n_i} \sqrt{n_j}} \left(n_{ij} - \frac{n_i n_j}{n} \right)$. Here, however, to keep the procedure as close as possible to Professor McQuitty's, we have worked with unstandardized values, and then applied the factorial procedure direct.

Professor McQuitty himself mentions certain analogies between agreement analysis and factor analysis. But he nowhere explains his objection to using factor analysis for his particular problem. It would, however, seem that, having American factorial procedures in mind, he regards it as restricted to 'linear continua' and incapable of providing a systematic classification into patterns when the data are discrete or qualitative instead of continuous or quantitative. British factorists commonly hold that factor analysis is itself primarily a classificatory device; and the method of 'subdivided factors' was expressly developed to provide a classification of the type Professor McQuitty requires, namely, one which will classify the persons or traits investigated into a ramifying scheme of genera, species, sub-species, and so on, each type being specified by its own distinctive pattern (see, for example, this *Journal*, II, pp. 41-63, esp. p. 46). The main difference is that 'agreement analysis' begins with the narrowest classes first, and then proceeds to the broader, whereas the factorial method starts by determining the widest class first of all, and then deals with the narrower classes in descending order, on the ground that it is wiser to begin with those components that account for the greatest amount of variance and therefore possess the greatest statistical significance.

Results. In applying the factorial method to Professor McQuitty's data, we confined ourselves mainly to the two simplest working procedures.

We began with the 'method of simple summation' (the so-called 'centroid procedure'), partly because it is the best known, partly because it provides the quickest way to determine the lines of demarcation between the various classes and subclasses. The factor loadings for the 'general' and 'bipolar' factors are shown in Table I A.

TABLE I. FACTOR ANALYSIS OF RESPONSES

Factors	A. Bipolar Factors			B. Group Factors				
	I	II	III	I	IIP	IIN	IIIP	IIIN
Persons								
<i>a</i>	5.56 +	.85 +	.55	4.93	2.46	—	1.35	—
<i>b</i>	4.61 +	.82 +	1.32	4.11	1.22	—	2.67	—
<i>f</i>	4.45 +	.79 +	.93	3.92	1.78	—	.88	—
<i>h</i>	4.98 +	.40 —	.68	4.40	2.30	—	—	1.95
<i>g</i>	4.89 +	.13 —	1.25	4.36	1.97	—	—	1.08
<i>c</i>	4.57 +	.71 —	.87	4.45	1.25	—	—	.61
<i>i</i>	4.20 —	1.84 +	.98	4.42	—	1.29	—	—
<i>d</i>	4.37 —	1.27 —	.47	4.34	—	1.29	—	—
<i>e</i>	3.86 —	.59 —	.51	3.97	—	—	—	—
Contribution to Variance	193.2	8.1	7.6	168.9	21.4	3.3	9.7	5.3
Per cent	59.6	2.5	2.3	52.1	6.6	1.0	3.0	1.6

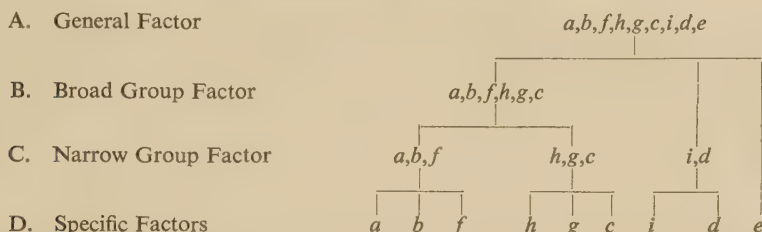
Note: The headings 'IIP' and 'IIN' indicate that the group factors so designated correspond to the positive and negative loadings for Factor II in the bipolar analysis; and similarly with IIIP and IIIN.

We then proceeded to carry out a group factor analysis, employing 'subdivided factors' for the largest group.¹ The loadings for the group factors so obtained are shown in Table I B.

¹ The method of rotation used was that described in this *Journal*, III, 1950, p. 75. In an earlier investigation into methods of discriminating potential delinquents and potential neurotics Burt based his rotation on a 'factorization of frequencies of a higher order' (see his Maudsley Lecture on 'The Assessment of Personality,' 1951). This procedure, though differently derived, appears to have a close analogy with Lazarsfeld's method of 'latent structure' referred to by Professor McQuitty (p. 6). For a brief exposition of the method see *J. Mental Science*, C, pp. 15 f.

Adopting the 'family tree' used by Burt to exhibit his hierarchical schemes of classification (e.g., this *Journal*, VII, p. 17), we can set out the essential conclusions as follows:

FIG. 1. *Hierarchical Classification of Persons*



The general or 'basic' factor corresponds with Professor McQuitty's 'single family F_1 '. The broad group factor, which includes a, b, f, h, g and c , corresponds with his 'genus', defined by the pattern GP_1 as shown in his Table II: the only difference is that he also includes e : (in Table III, f is dropped). The two narrower group factors correspond with his two 'species', defined by the patterns SP_1 and SP_2 ; but he classifies f with c , and a and b with h and g . Our other narrow factor, containing d and i , corresponds exactly with his third species SP_3 . In our analysis too e has a large specific factor; and that is evidently in accord with his impression. On the whole, then, the results of the factorial analysis correspond pretty closely with those of the 'agreement analysis'. A few minor discrepancies were to be expected since the sampling errors are so large.

The Relative Advantages of the Two Procedures. The factorial method seems to have one or two special advantages. To begin with, it renders the results directly comparable with those of other investigations. Our factor loadings, for example, like Professor McQuitty's 'agreement scores', express the extent to which each person is representative of this or that particular type; and we have only to divide these loadings by the square root of the variance for each person, namely 6, and each of the figures is then transformed into an ordinary product-moment correlation, i.e., into a 'factor saturation'. Indeed, in most investigations the covariances would have been standardized from the beginning, and so expressed in correlational form throughout.

Secondly, the factorial method enables us to determine the amount contributed by each type-factor to the total variance, i.e., the relative importance or salience of each type, and thus to check the statistical significance of each type-factor. With the present data, an unexpectedly large proportion of the total variance is contributed by the general factor (over 50 per cent. with the group factor analysis and nearly 60 per cent. with the bipolar analysis). On the other hand, the amount contributed by the type-factors is decidedly small: little over 6 per cent. is contributed by the broad group factor, and a rather smaller amount is divided among the three narrow factors. Indeed, as the bipolar analysis plainly indicates, the statistical significance of the type factors is extremely doubtful. As already suggested, it is this that probably accounts for the slight divergences between the two classifications.

In the third place, the factorial procedure enables us to determine the 'factor measurements' for each item or trait: when, as in the present case, the method of correlating persons is adopted, these 'factor measurements for items' correspond with the so-called 'factor loadings for items' which appear when the method of correlating test-items is adopted. From these 'factor measurements' a psychogram or profile can be constructed exhibiting in graphic form the pattern of responses characteristic of each type (cf., *Factors of the Mind*, Figure 1, p. 427). Professor McQuitty also specifies the pattern of responses for the several types (Table I); but his entries are limited to 'Y' or 'N', i.e., to mere presence or absence, whereas our factor measurements not only indicate whether the diagnostic implication of each item is positive or negative, but also express its importance by a coefficient of varying magnitude, similar to a coefficient of correlation.

We could thus rearrange the initial data (Table I in the preceding article) so as to bring out in visible form the detailed classification both of persons and of items. The cost of

printing will not permit us to reproduce our own rearrangement in full; but the reader will find it most illuminating to attempt one for himself, either with or without the aid of a prior factor analysis.¹ With dichotomous data like the present, the underlying principle is to grade both columns and rows so that the resulting scheme approximates as closely as possible to what is loosely termed a 'triangular answer-pattern', i.e. one in which positive entries and negative entries (or units and zeros) are each brought as near together as possible, and the scheme as a whole presents a more or less regular step-like descent: (see, for example, the theoretical 'answer-pattern' set out in this *Journal*, VI, p. 7, Table I).

With the present data, this procedure yields, first, a large block of items at the top (4, 14, 15, 17, 21, 33, 35) to which practically every person answers 'Yes', and secondly, another block at the bottom (2, 12, 23) to which everyone answers 'No' or 'Doubtful', and we may add another half-dozen to which all except one return such answers. It is these two large blocks that chiefly account for the unusual size of the general factor. Their presence implies that out of the whole list of items there are really only about 20 which can provide information for the classification by types.

Of the various types and sub-types the broadest of all (that containing *a*, *b*, *f*, *g*, *c* and *h*) is distinguished most clearly by items 7, 9, 16 and 20, to which the members of the type usually answer 'No', while the rest generally say 'Yes' or 'Doubtful', and by items 8, 10, 18 and 22, to which the type-members give either a definite 'Yes' or a definite 'No', while the rest say 'Doubtful'. This broad type is subdivided by items like 1, 6 and 25, to which *a*, *b* and *f* tend to say 'Yes', while *g*, *c* and *h* tend to say 'No' or 'Doubtful'. Of those who, like *d*, *i* and *e*, are not members of the broadest type, *d* and *i* are further distinguished from *e*, first, by the fact that they answer 'Yes' to 7, 8, 19, 20 and 29, while *e* says 'No' or 'Doubtful', and second, by the fact that they answer 'No' to 24, 27, 28 and 34, while *e* says 'Yes' or 'Doubtful'. As we have seen, both *i* and *e* have large specific factors: *i* tends to say 'Yes' to nearly all the items not in the 'Yes' or 'No' blocks, whereas *e* tends nearly always to say 'Doubtful'; in fact, he says 'Doubtful' to pretty well one-third of all the items.

One last conclusion emerges from our factorial study which should, we think, be of considerable practical utility: namely, if the main purpose of the questionnaire is to classify persons into types, this function could be performed just as efficiently if one half of the items were omitted, since 18 out of the 36 have non-significant 'factor-loadings'.

¹ To our mind, it is one of the merits of Burt's technique that he insists on the value of *significant order* in arranging the matrices obtained in factorial work, whether they consist of raw scores, observed correlations, or residual coefficients. The items in the tables are regrouped according to their structural relations. Hence, if there are any causal regular ties underlying the empirical data, they frequently become manifest at a glance.

THE MOST RELIABLE COMPOSITE WITH A SPECIFIED TRUE SCORE¹

By MAX A. WOODBURY and FREDERIC M. LORD
George Washington University · Educational Testing Service

Summary. If a composite score is to be obtained from a battery of tests, the manner in which these tests enter into the composite may be varied in two respects: (a) different weights may be assigned to the tests, and (b) the test lengths (and correspondingly the administration times) may be altered. Weights and administration times may be assigned so as to maximize either the validity or the reliability of the composite. The present paper comments briefly on the problem of maximum validity, and then presents a simple method for maximizing reliability.

I. MAXIMIZING VALIDITY

In cases where an outside criterion is available, the problem of finding weights and administration times that will maximize the validity of the composite has been considered by Horst [2, 3] and by Taylor [5]. However, there are two serious difficulties, not generally appreciated, in the practical application of the formulae there given. First, they may give a solution involving negative administration times. In such a case, it is *not* sufficient simply to omit the tests for which the administration times are found to be negative, since there is no reason to believe that the remaining tests are necessarily those that belong in the optimum battery. Secondly, both Horst's and Taylor's solutions depend on formulae involving square roots. If the positive square root is taken in Horst's formulae, these can only provide the solution for the given matrix of test intercorrelations, the signs of the correlations being fixed. The practical worker, however, is quite willing to reflect the scores on one or more tests (i.e., to change the signs from positive to negative) if this will improve the validity of the battery as a whole. Taylor meets this problem by explicitly attaching a plus-or-minus symbol to his square root signs, without, however, supplying any simple rule for determining which sign to use. The practical worker has no alternative but to try out all possible combinations of signs for the n tests. If no optimum battery with non-negative administration times is found for n tests, all possible sets of $n-1$ tests must next be investigated. There are n such sets, and each must be investigated for each of the 2^{n-2} possible assignments of positive and negative scores to determine which set and which assignment give the maximum validity while avoiding negative administration times. If all sets of $n-1$ tests lead to negative administration times when Horst's or Taylor's formulae are applied, then all possible sets of $n-2$ tests must be tried out, and so on.

II. MAXIMIZING RELIABILITY

If no outside criterion is available and if the administration time of each test is fixed, the weights that will maximize the reliability of the composite may be determined by the method of canonical correlation, as applied to this problem by Thomson [6] and by Peel [4] (summarized in [1], p. 346-348). The present paper deals with the problem of maximizing the reliability of the composite when both the weight and the administration time of each test can be varied.

If no restrictions are placed on the composite, the solution becomes trivial, since in

¹ This project has been supported by the Office of Naval Research under Contract N6onr-270-20 and by the National Science Foundation, Grant NSF G-642. The basic mathematical derivation is the work of the senior author. The authors are indebted to Professor Harold Gulliksen for his encouragement and penetrating comments on a draft of the manuscript.

the optimum composite all the available administration time will be given to the one test in the battery that has the greatest reliability at unit length (or the shortest 'standard length' as defined in [7]). An optimum solution of practical value can be obtained by specifying the true score on the composite and holding this fixed while the reliability is maximized.

In dealing with the problems thus outlined, it will be assumed that the true scores on the different tests in the battery are linearly independent, i.e., that no true score is exactly equal to a weighted average of the other true scores. For this a necessary and sufficient condition is that the matrix of test inter-correlations be non-singular when the test reliability coefficients are placed in the diagonal. The rather extraordinary result will be obtained that, *except for sign, the weight allotted to any test in the optimum composite with a predetermined true score depends only on the standard error of measurement of that test at unit length.* This optimum weight (except for sign) does not depend on the true score of the composite, nor on the nature of the tests included in the battery, nor on the total testing time available.

Simple formulae for determining the optimum weight and the optimum testing time for every test in the battery will be presented. The optimum scoring weight for each test will be shown to be simply the reciprocal of the standard error of measurement at unit length. Gulliksen, it may be noted, has already suggested the reciprocal of the standard error of measurement at actual (rather than unit) length as a weight suitable for general use (1, p. 336).

III. MATHEMATICAL STATEMENT OF PROBLEM

Let t_i denote the (variable) amount of administration time devoted to items of type i . It will be assumed that the number of items of a given type administered per unit of time always remains the same. Hence, for any given test, t_i may be considered proportional to the number of items of type i administered, and may be referred to as the 'length' of test i .

Let $x_i(t_i)$ be the score on test i (i.e., on the items of type i) when the time allowed is t_i . The usual assumption will be made that $x_i(t_i)$ behaves as if composed of a true-score component $\xi_i(t_i)$ plus an independent error component $\epsilon_i(t_i)$ having zero mean: i.e.,

$$x_i(t_i) = \xi_i(t_i) + \epsilon_i(t_i), \quad \text{and} \quad x_i(1) = \xi_i + \epsilon_i, \quad (1)$$

where ξ_i is the true score of test i at unit length and ϵ_i is the error component at unit length.

It is assumed that the proportion of items that an examinee answers correctly is independent of the number of items in the test. Hence ξ_i may be defined by the equation

$$\xi_i = \lim_{t_i \rightarrow \infty} \frac{t_i}{1} x_i(t_i), \quad (2)$$

and the error ϵ_i by the equation

$$\epsilon_i = x_i(1) - \xi_i. \quad (3)$$

It follows that the mean score of an infinite population of examinees is (cf. 1, chap. ii, eqn. (9))

$$M_{x_i(t_i)} = t_i \bar{\xi}_i; \quad M_{x_i(1)} = \bar{\xi}_i, \quad (4)$$

where $\bar{\xi}_i$ is the mean true score of all examinees; and the variance over an infinite population of examinees is (cf. 1, chap. ii, eqn. (17))

$$\sigma_{x_i(t_i)}^2 = t_i^2 \sigma_{\xi_i}^2 + t_i \sigma_{\epsilon_i}^2, \quad \sigma_{x_i(1)}^2 = \sigma_{\xi_i}^2 + \sigma_{\epsilon_i}^2. \quad (5)$$

Let us define reliability as the ratio of true-score variance to total-score variance (cf. 1, chap. x, eqn. (8)): so that

$$r_{ii}(t_i) = \frac{t_i \sigma_{\xi_i}^2}{t_i \sigma_{\xi_i}^2 + \sigma_{\epsilon_i}^2}. \quad (6)$$

Finally, let us write down the covariance between test i and test j :

$$\sigma_{x_i(t_i), x_j(t_j)} = t_i t_j \sigma_{ij} + \delta_{ij} t_i \sigma_{\epsilon_i}^2, \quad (7)$$

where σ_{ij} is the covariance between ξ_i and ξ_j in the population of examinees, and $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$.

Now consider the weighted composite score

$$y = \sum_{i=1}^n w_i x_i(t_i). \quad (8)$$

The true score (η) of this composite at unit time is proportional to the corresponding composite of true scores, i.e.,

$$\eta \propto \sum_i w_i t_i \xi_i. \quad (9)$$

The variance of y , by eqn. (7) and the usual formula for the variance of a sum, will be

$$\sigma_y^2 = \sum_i \sum_j t_i t_j w_i w_j \sigma_{ij} + \sum_i w_i^2 t_i \sigma_{\xi_i}^2. \quad (10)$$

The reliability of the composite is thus

$$r_{yy} = \frac{\sum_i \sum_j w_i t_i w_j t_j \sigma_{ij}}{\sum_i \sum_j w_i t_i w_j t_j \sigma_{ij} + \sum_i w_i^2 t_i \sigma_{\xi_i}^2}. \quad (11)$$

The problem is to maximize r_{yy} subject to the following restrictions:

1. The total administration time $\sum_i t_i$ is limited to a maximum of t , i.e.,

$$\sum_i t_i \leq t. \quad (12)$$

2. The testing time for each test must be non-negative, i.e.,

$$t_i \geq 0 \quad (i = 1, \dots, n). \quad (13)$$

3. The true score of the composite is specified except for an unknown coefficient of proportionality, i.e.,

$$\eta \propto \sum_i w_i t_i \xi_i = p \sum_i \gamma_i \xi_i, \quad (14)$$

where the γ_i are given constants, and p is an unspecified coefficient of proportionality. Subject to these restrictions, r_{yy} is to be maximized by a choice of values for w_i , t_i , and p .

The values of γ_i (except for an arbitrary and irrelevant coefficient of proportionality) must always be specified in advance by the testing expert on the basis of subjective judgement or special information, just as the types of tests (vocabulary, arithmetic, etc.) are themselves chosen by the testing expert. He need only ask himself the question: if an unweighted composite were to be used, what should be the administration time assigned to each test for my present purposes? If he can answer this question, then, by (19'), the administration times that he chooses may be treated as the desired values of γ_i . Once the values of γ_i are given, the formulae that will be presented can be used to obtain the most reliable composite having the same true score as that specified by the testing expert.

In practice it may frequently be desired to determine if the composite score on a given battery in current use can be made more reliable without changing its true score. What are the values of γ_i corresponding to the true score of the composite in current use? If each test has administration time T_i and receives equal weight in the current composite, then by (19') the required values of γ_i are proportional to the T_i , the coefficient of proportionality being immaterial. Starting with these values of γ_i , it is easy to determine if the reliability of the composite can be increased.

In summary, the following will be proved. Given

- (1) n tests with scores represented by $x_i(T_i)$ ($i = 1, \dots, n$), with known
 - (i) administration time T_i ,
 - (ii) test reliability, $r_{ii}(T_i)$,
 - (iii) variance-covariance matrix, $\|\sigma_{x_i(T_i), x_j(T_j)}\|$,
- (2) total available testing time, t ,

Most Reliable Composite with Specified True Score

(3) n constants γ_i ($i = 1, \dots, n$) defining the true score of the desired composite, then the most reliable composite with specified true score will be obtained by the following procedure (assuming always that the true scores on the n tests are not linearly dependent):

- (1) compute σ_{ϵ_i} for each test from eqn. (27);
- (2) compute $\hat{t}_i$ for each test from eqn. (26);
- (3) administer the tests, allotting an administration time of $\hat{t}_i$ to test i ;
- (4) compute $\hat{w}_i$ for each test from eqn. (25);
- (5) compute the optimum composite, using a weight of $\hat{w}_i$ for test i .

The reliability of the optimum composite may be computed from eqn. (11) and (28) if t_i is substituted for t_i and T_i in these formulae and $\hat{w}_i$ for w_i .

IV. PRELIMINARY THEOREMS

Since restrictions (12) and (13) involve some inequalities, a direct approach using Lagrange multipliers to maximize (11) encounters difficulties. It will be more satisfactory to work with the following theorems.

Theorem 1. *If $t, t_1, t_2, \dots, t_n, w_1, w_2, \dots, w_n$ are real numbers, with each $t_i > 0$, such that (i) $\sum_i w_i^2 t_i < t$, and (ii) $\sum_i t_i < t$, then, (iii) $\sum_i |c_i| < t$, where (iv) $c_i = w_i t_i$.*

This is proved by writing the quantity on the left of (iii) as $\sum_i |w_i \sqrt{t_i}| \sqrt{t_i}$. By the Schwartz inequality,¹

$$\sum_i |w_i \sqrt{t_i}| \sqrt{t_i} < \sqrt{(\sum_i w_i^2 t_i)} \sqrt{\sum_i t_i}. \quad (15)$$

By (i) and (ii), each of the two radicals on the right of (15) is equal to or less than $\sqrt{t}$. We have

$$\sum_i |c_i| = \sum_i |w_i \sqrt{t_i}| \sqrt{t_i} < \sqrt{(\sum_i w_i^2 t_i)} \sqrt{\sum_i t_i} < t. \quad (16)$$

Hence eqn. (iii).

Theorem 2. *If $c_i \neq 0$ ($i = 1, 2, \dots, n$) are real numbers such that (v) $\sum_i |c_i| = t$, then $t_i > 0$ and real w_i satisfying (i), (ii), and (iv) will be uniquely determined by the equations:*

(vi a) $w_i = \frac{c_i}{|c_i|}$, and (vi b) $t_i = |c_i|$; furthermore, (vii) $\sum_i t_i = t$, and (viii) $\sum_i w_i^2 t_i = t$.

To prove Theorem 2, note that, in view of (v), (16) becomes $t \leq \sqrt{(\sum_i w_i^2 t_i)} \sqrt{\sum_i t_i} \leq t$.

Consequently, only the equality sign can hold in (ii) and (iii), thus proving eqn. (vii) and (viii). Further, because of the Schwartz inequality, we know that the equality sign in (15) holds only if $|w_i \sqrt{t_i}|$ is proportional to $\sqrt{t_i}$; and consequently $|w_i|$ must be a constant, since $c_i = w_i t_i \neq 0$. From (v), (vii), and (viii) it immediately follows that the only possible values for w_i and t_i are those given by (vi). The notation in (vi) indicates that $w_i = +1$ or $w_i = -1$ depending on the sign of c_i .

V. FORMULAE FOR MAXIMIZING RELIABILITY

Any choice of unit of measurement may be made for each test score without affecting the test reliability. The choice will be made so that

$$\sigma_{\epsilon_i} = 1 \quad (i = 1, 2, \dots, n). \quad (17)$$

Primes will be used to indicate statistics based on this particular scale of measurement.

It has been seen from eqn. (11) that the battery reliability is a homogeneous function of the w_i (i.e., the reliability is unchanged if all w_i are multiplied by a constant). Hence, the w_i are indeterminate to the extent of a constant factor, which may be chosen arbitrarily. It will be convenient to choose this constant so that the error-score variance of the composite is equal to t , i.e., so that

$$\sum_i t_i w_i'^2 = t. \quad (18')$$

¹ If u_i and v_i ($i = 1, \dots, n$) are any real numbers, then $(\sum u_i v_i)^2 \leq (\sum u_i^2)(\sum v_i^2)$; the equality holds if and only if u_i is proportional to v_i .

Since (14) is an identity that holds for each examinee separately, the coefficients of ξ_i can be identified for each value of i , so that

$$p\gamma_i = w_i t_i, \quad \text{and} \quad p\gamma'_i = w'_i t_i \quad (i = 1, 2, \dots, n) \quad (19')$$

(provided as before that the ξ_i are not linearly dependent). The value of p must be chosen so as to maximize the battery reliability. Under conditions (17) and (18'), it is immediately seen that the battery reliability will be maximized provided merely that the numerator of eqn. (11) is maximized. By (19'), this numerator, which will be denoted by r' , is

$$r' = p^2 \sum_i \sum_j \gamma'_i \gamma'_j \sigma'_{ij}. \quad (20')$$

Now γ'_i , γ'_j , and σ'_{ij} are all given. Consequently, r will be maximized by choosing p as large as possible, subject to the restrictions imposed by eqn. (12), (13), (18'), and (19').

From eqn. (19') and from (iv) of Theorem 1, it is seen that

$$p\gamma'_i = c'_i. \quad (21')$$

From (21') and (iii),

$$\sum_i |p\gamma'_i| < t. \quad (22')$$

Since we choose p as large as possible, let us see if we can choose p so that the equality holds in (22') without violating restrictions (12), (13), and (18'). If so, this is the value of p required to solve our problem. (The optimum value of a parameter will be denoted by placing a circumflex accent or 'hat' over the symbol.)

If p is chosen so that $\sum_i |\hat{p}\gamma'_i| = t$, then, by (19') and (iv), eqn. (v) of Theorem 2 will hold.

This guarantees that the w'_i and non-negative t'_i defined by (vi) exist, are unique, and satisfy eqn. (vii), thus meeting the conditions imposed by restrictions (12) and (13), and also satisfy eqn. (viii), thus meeting the condition imposed by (18'). Hence

$$\hat{p} = t / \sum_i |\gamma'_i| \quad (23')$$

is the desired value of p .

It follows from (19') and (iv) that

$$\hat{c}'_i = \gamma'_i t / \sum_i |\gamma'_i|, \quad (24')$$

i.e., the $\hat{c}'_i$ are proportional to the γ'_i , the coefficient of proportionality being so chosen that $\sum_i |\hat{c}'_i| = t$. From (vi),

$$\hat{w}'_i = \hat{c}'_i / |\hat{c}'_i| = \gamma'_i / |\gamma'_i|, \quad (25')$$

i.e., $\hat{w}'_i$ is either +1 or -1 depending on the sign of $\hat{c}'_i$. From (vi),

$$\hat{t}_i = |\hat{c}'_i|. \quad (26')$$

Eqn. (24'), (25') and (26') constitute the solution to the problem of finding the optimum weights and optimum administration times in the special case where all scores are scaled so that $\sigma_{\epsilon_i} = 1$. It is seen from (24'), (25') and (26') that if (a) *all tests receive equal weight in forming the composite*, (b) *all tests have equal standard errors of measurement at unit time*, then the reliability of the composite cannot be increased by weighting the tests and reshuffling the test lengths without altering the composite true score. Incidentally, it is apparent that the optimum composite will have the greatest advantage over the usual unweighted composite in the case where the individual tests have distinctly different standard errors of measurement at unit time.

Supplementary Formulae for Computational Purposes. It has been assumed that a single scoring weight (here considered distinct from the weight w_i applied to the test score in producing the weighted composite) has been used for all the items in a single test, the weight being so chosen that $\sigma_{\epsilon_i} = 1$. This scoring weight was introduced to display the simplicity of the results obtained, and to indicate the possibility of incorporating such weights into the

Most Reliable Composite with Specified True Score

scaled-score system of any and all tests. The value of σ_{ε_i} is readily derived from eqn. (5) and (6), viz.,

$$\sigma_{\varepsilon_i} = \frac{1}{\sqrt{T_i}} \sigma_{x_i(T_i)} \sqrt{1 - r_{ii}(T_i)}, \quad (25)$$

where T_i is the length of the actual test whose reliability and standard deviation are known. The required scoring weight will be the reciprocal of (25).

It will be assumed throughout the remainder of this article that no such scoring weight has been applied. If it is not applied during the scoring, it must be included in the w_i , i.e., in the weighting used to obtain the composite score. In this case,

$$\hat{w}_i = \frac{\gamma_i}{|\gamma_i|} \frac{1}{\sigma_{\varepsilon_i}}. \quad (26)$$

Since, by (19'), $\hat{t}_i \propto \gamma_i / \hat{w}_i$, we have $t_i = |\gamma_i| \sigma_{\varepsilon_i}$; and since $\sum \hat{t}_i = t$, it follows that now

$$\hat{t}_i = \frac{t}{\sum_i |\gamma_i| \sigma_{\varepsilon_i}} |\gamma_i| \sigma_{\varepsilon_i}. \quad (27)$$

In practice, one may wish to know the numerical value of the battery reliability obtained by use of the optimum weights and administration times. This can be readily ascertained from eqn. (11). Computational formulae have been presented for all the symbols involved in eqn. (11) except σ_{ij} , the covariance of ξ_i and ξ_j . From (7) and (25),

$$\sigma_{ij} = \begin{cases} \frac{1}{T_i T_j} \sigma_{x_i(T_i) x_j(T_j)} & \text{if } i \neq j \\ \frac{r_{ii}(T_i)}{T_i^2} \sigma_{\varepsilon_i}^2 & \text{if } i = j \end{cases} \quad (28)$$

Numerical Examples. Fig. 1 illustrates how composite-score reliability varies under various circumstances. The figure relates to a battery composed of two hypothetical tests. At unit time ($T_1 = 1$, $T_2 = 1$), the test statistics for these two tests are $\sigma_{x_1(1)} = 1$, $\sigma_{x_2(1)} = 1$, $r_{11}(1) = .90$, $r_{12}(1) = .70$, $r_{22}(1) = .60$. For each of five different true scores, as specified by the values of γ_1 and γ_2 , the figure shows the reliability $r_{vv}(1)$ of the composite of the two tests as a function of the proportion (t_1) of total testing time allocated to test 1. The total testing time is here taken as 1, so that $t = t_1 + t_2 = 1$. For, given γ_1 and γ_2 , each value of t_1 determines corresponding values of w_1 and w_2 : see eqn. (19'). These values were substituted in (11) together with values of σ_{ij} and σ_{ε_i} obtained from the given data by means of eqn. (28) and (25), in order to plot the curves in Fig. 1.

The optimum values of w_1 and w_2 are found from (26) to be 3.16 and 1.58. These optimum values, of course, are independent of the composition of the composite true score. It will be seen from the figure that, when the true composite score involves the true scores on tests 1 and 2 in the ratio of 4 to 1 ($\gamma_1 = .80$, $\gamma_2 = .20$), the maximum reliability is attained when $t_1 = \hat{t}_1 = .667$, approximately. This same value may be computed by eqn. (25) and (27) from the data given. The unweighted composite having the same true score is obtained (see eqn. (19)) when $t_1 = \gamma_1 = .80$ and $t_2 = \gamma_2 = .20$. The reliability of the unweighted composite is found to be $r_{vv} = .837$. This value is numerically only slightly less than that of the optimally weighted composite having the same true score ($r_{vv} = .851$); however, the length of a test would have to be increased 11 per cent. in order to increase its reliability from .837 to .851. The reliabilities of other unweighted and optimally weighted composites are compared in Table I.

VI. SUMMARY AND CONCLUSIONS

Formulae have been derived for weighting the tests in a battery and allocating the administration time so as to maximize the reliability of the weighted battery composite, the

true score of the composite being specified in advance. For the purpose of this derivation the true scores on the different tests in the battery were assumed to be linearly independent.

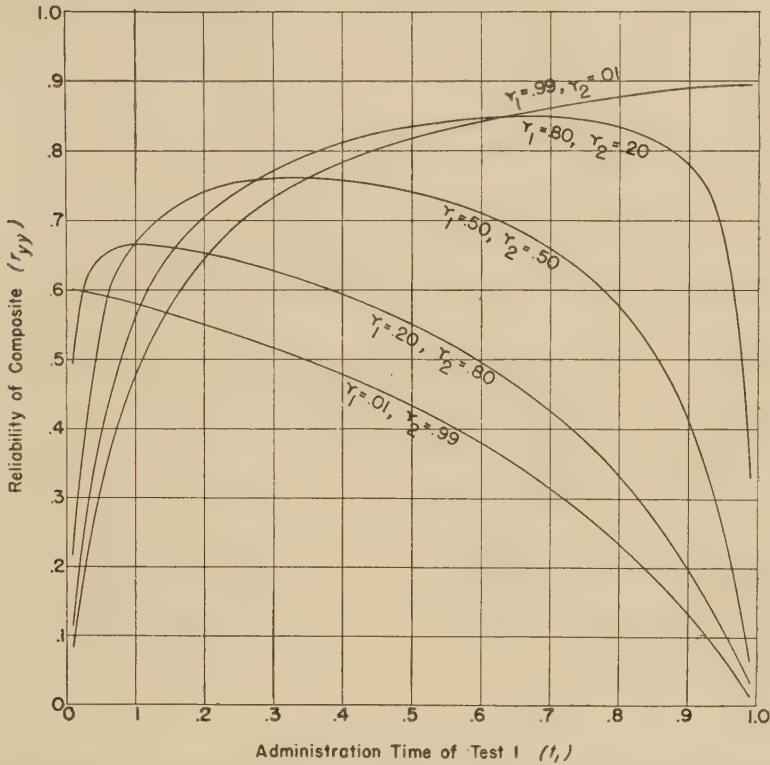


FIG. 1.—The reliability of all possible composites, for each of five different true scores, shown as a function of the allocation of testing time.

TABLE I. RELIABILITY OF AN UNWEIGHTED COMPOSITE AND OF THE OPTIMALLY WEIGHTED COMPOSITE HAVING THE SAME TRUE SCORE

True Score		Unweighted Composite			Optimally Weighted Composite		
γ_1	γ_2	t_1	w_1	$r_{yy}(1)$	$\hat{t}_1$	$\hat{w}_1$	$r_{yy}(1)$
.20	.80	.20	1	.654	.111	3.2	.665
.50	.50	.50	1	.744	.333	3.2	.763
.80	.20	.80	1	.837	.667	3.2	.851

The proportion of the available testing time assigned to each test in the optimum battery is a function of the standard error of measurement of that test at unit time and of the desired composite true score. This proportion does not depend on total testing time available, nor on any statistics involving other tests in the battery, except in so far as these statistics

Most Reliable Composite with Specified True Score

influence composite true score. Except for sign, the weight assigned to any test in the optimum composite is determined solely by the standard error of measurement of that test at unit time.

The foregoing results are especially appropriate in the many situations in achievement testing where there is no better criterion than an expertly devised achievement test. The expert devising such a test should first specify in his own mind the particular kinds of knowledge and skill to be measured; this corresponds to specifying the true score desired for the composite. He should then attempt to build a battery that will yield scores having the highest possible correlation with the specified true score; this is the same as building a battery measuring the desired true score and having the maximum possible reliability. To achieve this goal, he need only weight each test item in scoring by the reciprocal of the standard error of measurement of a unit-length test composed of such items. Once this is done, whatever battery the expert may build automatically will have a higher reliability than any other battery composed of the same kinds of items and having the same true score.

In the foregoing situations, where the true score to be measured is specified by the expert, maximizing the reliability with which this true score is measured actually maximizes validity. In fact, it is the only way to maximize validity in such situations.

REFERENCES

1. Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley and Sons.
2. Horst, P. (1949). 'Determination of optimal test length to maximize the multiple correlation.' *Psychometrika*, XIV, 79-88.
3. Horst, P. (1950). 'A note on optimal test length.' *Psychometrika*, XV, 407-408.
4. Peel, E. A. (1948). 'Prediction of a complex criterion and battery reliability.' *Brit. J. Psychol., Statist. Sect.*, I, 84-94.
5. Taylor, C. (1950). 'Maximizing predictive efficiency for a fixed total testing time.' *Psychometrika*, XV, 391-406.
6. Thomson, G. H. (1940). 'Weighting for battery reliability and prediction.' *Brit. J. Psychol.*, XXX, 357-366.
7. Woodbury, M. (1951). 'On the standard length of a test.' *Psychometrika*, XVI, 103-106

PREDICTION OF SCALE VALUES FOR COMBINED STIMULI

By H. J. A. RIMOLDI¹

Loyola University, Chicago, Illinois

I. *Introduction.* II. *Determination of Theoretical Values.* III. *Reproduction of Original Proportions.* IV. *Experimental Procedure.* V. *Experimental Results.* VI. *Discussion.*

I. INTRODUCTION

This study investigates the relation between the scale value of single stimuli and the scale value of their combinations when given in groups of two. The scale value for combined stimuli has been considered to be a new experience (Wundt's *Totalgefühl*), or the algebraic sum of their components (Titchener and McDougall), and so forth. J. P. Guilford and W. Spence [3] and J. P. Guilford [2] studied the problem experimentally in relation to colours and to odours, and concluded that the 'affective' reaction to the combination of single stimuli is the weighted mean of the affect belonging to these single stimuli. In these studies no attempt was made to establish the analytical procedure relating the scale value corresponding to the double stimuli with the scale value pertaining to the single stimuli.

The main steps to be followed in the present study are as follows:

- (1) Determine experimentally the discriminial processes and discriminial dispersions corresponding to the single stimuli.
- (2) Determine experimentally the discriminial process and discriminial dispersion corresponding to the double stimuli, that is, the combination of single stimuli in groups of two.
- (3) From (1) above determine analytically the theoretical (predicted) values for the discriminial processes and the discriminial dispersions corresponding to the double stimuli.
- (4) Compare the results obtained in (2) and (3) above.

II. DETERMINATION OF THE THEORETICAL VALUES

The notation used was the following:

S_j and σ_j represent respectively the experimentally obtained modal discriminial process and discriminial dispersion for stimulus j , where $j = 1, 2, \dots, j, \dots, n$.

S_{jk} and σ_{jk} represent respectively the experimentally obtained modal discriminial process and discriminial dispersion for stimulus jk , where $jk = 12, 13, \dots, jk, \dots, (n-1)n$ and $j \neq k$.

S_{jk}^* and σ_{jk}^* represent the theoretical values (predicted) for the modal discriminial process and the discriminial dispersion for the double stimuli.

r_{jk} is the correlation between the subject's judgement of stimuli j and k .

L_i ($i = 1, 2, \dots, i, \dots, m-1$) represents the modal discriminial process of the boundary between intervals i and $i+1$, where there are m successive intervals. p_{ij} , p_{ik} , p_{ijk} , p_{ijk}^* represent the proportion of times that a stimulus has been judged to be smaller than i , and X_{ij} , X_{ik} , X_{ijk} and X_{ijk}^* are their corresponding normal deviates.

Two hypotheses were tested in this study:

- (a) Hypothesis of linearity as indicated in (1)

$$S_{jk}^* = \frac{S_j + S_k}{2} \quad (1)$$

¹ This study was conducted during the years 1951 and 1952 at the 'Universidad de la República', Montevideo, Uruguay. The author wishes to thank Miss M. Hormaeche for her invaluable help in performing the experiment and most of the calculations.

Prediction of Scale Values for Combined Stimuli

(b) The relation between the affective value for the double and the single stimuli is better explained by assuming the law of diminishing returns.¹

It will be assumed that the double and the single stimuli are normally distributed on the same psychological continuum.

We shall define:

$\sum S_j = 0$ (2), $\sum S_{jk} = 0$ (3), $\sum S_{jk}^* = 0$ (4), $\sum \sigma_j = n$ (5), $\sum \sigma_{jk} = n'$ (6), where $n' = \frac{n(n-1)}{2}$, and $\sum \sigma_{jk}^* = n'$ (7). Eqn. (2) and (5) represent respectively the summation of all the S_j and σ_j values, and eqn. (3), (4), (6), and (7) the summation of all the S_{jk} , S_{jk}^* , σ_{jk} and σ_{jk}^* values.

$$\text{Let } X_{ijk}\sigma_{ik} = [X_{ijk}^*\sigma_{jk}^*]e + K \quad (8)$$

where e is proportional to the ratio of the dispersions between the $X_{ijk}\sigma_{jk}$ and the $X_{ijk}^*\sigma_{jk}^*$ values, and K is an additive constant to take into consideration differences in origin for the two sets of values.

Adding eqn. (8) for all the i values and for all the jk values, we obtain respectively eqn. (9) and (10)

$$\sigma_{jk}\sum X_{ijk} = e\sigma_{jk}^*\sum X_{ijk}^* + (m-1)K \quad (9)$$

$$\text{and } \sum X_{ijk}\sigma_{jk} = e\sum X_{ijk}^*\sigma_{jk}^* + n'K. \quad (10)$$

In a previous paper [5] it was demonstrated that, for single stimuli $S_j = \frac{\sum L_i - \sigma_j \sum X_{ij}}{m-1}$ (11), where the summation sign extends over all the i values and $L_i = \frac{\sum X_{ij}\sigma_j}{n}$ (12), where the summation sign indicates adding for all the j values.

Accordingly, adding all the i values for the double stimuli, we can write:

$$S_{jk} = \frac{\sum L_i - \sigma_{jk} \sum X_{ijk}}{m-1} \quad (13), \quad \text{where } L_i = \frac{\sum X_{ijk}\sigma_{jk}}{n'} \quad (14).$$

The summation sign in eqn. (14) applies to all the jk values.

Hence, eqn. (10) becomes: $L_i = eL_i^* + K$ (15), where L_i^* represents the theoretically determined value for boundary i on the continuum. L_i^* is obtained from eqn. (14) by replacing $X_{ijk}\sigma_{jk}$ by $X_{ijk}^*\sigma_{jk}^*$.

Substituting (15) and (9) into (13), we obtain: $S_{jk} = \frac{e[\sum L_i^* - \sigma_{jk}^* \sum X_{ijk}^*]}{m-1} = eS_{jk}^*$ (16), where the summation sign refers to all the i values.

According to the basic scaling equation given in the paper just cited [5] we can write:

$$X_{ijk}^*\sigma_{jk}^* + S_{jk}^* = X_{ij}\sigma_j + S_j, \quad X_{ijk}^*\sigma_{jk}^* + S_{jk}^* = X_{ik}\sigma_k + S_k. \quad (17)$$

$$\text{Hence, } X_{ijk}^*\sigma_{jk}^* = \frac{X_{ij}\sigma_j + X_{ik}\sigma_k}{2} + \frac{S_j + S_k}{2} - S_{jk}^* \quad (18)$$

which according to (1) becomes:

$$X_{ijk}^*\sigma_{jk}^* = \frac{X_{ij}\sigma_j + X_{ik}\sigma_k}{2}. \quad (19)$$

¹ Early in 1953 the author discussed the method and the results given in this paper with H. O. Gulliksen, who suggested trying the law of diminishing returns. Our data could not be explained in terms of the law of diminishing returns. In 1954 in a personal communication with the author, Gulliksen reported that in a similar experiment, using the paired comparisons method, he had also verified the hypothesis of linearity as stated in the present article.

We define:
$$\sigma_{jk}^* = c[\frac{1}{2}\sqrt{(\sigma_j^2 + \sigma_k^2 + 2r_{jk}\sigma_j\sigma_k)}] \quad (20)$$

where c is a constant to make, according to definition (6), $\Sigma\sigma_{jk}^* = n'$, and where the summation sign extends over all the jk values. It can be demonstrated that

$$c = \frac{n'}{\Sigma[\frac{1}{2}\sqrt{(\sigma_j^2 + \sigma_k^2 + 2r_{jk}\sigma_j\sigma_k)}]} \quad (21)$$

III. REPRODUCTION OF THE ORIGINAL PROPORTIONS

The distance between two points L_i^* and S_k^* on the continuum can be written:

$$L_i^* - S_k^* = X_{ijk}^* \sigma_{jk}^* \quad (22)$$

Hence,
$$X_{ijk}^* = (L_i^* - S_k^*) \frac{1}{\sigma_{jk}^*} \quad (23)$$

Using a probability integral table, the p_{ijk}^* values corresponding to each X_{ijk}^* can be found. It is expected that the discrepancies between p_{ijk} and p_{ijk}^* will be a minimum. Eqn. (1) to (23) refer to true values. For the purposes of computation a parallel set of equations can be written, using sample values instead of true values.

IV. EXPERIMENTAL PROCEDURE

The experimental group consisted of 170 adult students enrolled in teacher training institutions in Montevideo, Uruguay. The tests were administered to smaller groups varying between 20 and 30 subjects. A more detailed account of the experimental procedure is given in [5]. The single and double stimuli consisted respectively of names of famous people given singly or in pairs. These names are given in Tables I, II, III, and IV. Both types of stimuli were presented in three different ways.

(a) One Graphic Rating Scale (G.R.S.) was used to rate 25 single stimuli and another to rate 21 double stimuli, generated by combining 7 of the single stimuli in groups of 2. The subjects were asked to mark on the continuum according to their interest 'in knowing' the person or persons represented by the stimuli. For the purpose of scoring the test, 9 arbitrary intervals were superimposed on the continuum.

(b) A similar procedure was followed when using the Method of Successive Intervals (M.S.I.) for 15 single stimuli and for 22 double stimuli. In both scales there were 9 successive intervals. The 22 double stimuli were obtained by combining 8 single stimuli in groups of 2. One of these single stimuli was used in only one combination.

(c) Multiple Category Method (M.C.M.) [5] for 15 single stimuli and 21 double stimuli, obtained by combining 7 single stimuli in groups of 2. The subjects had to encircle the word that best described their interest in knowing the person or persons indicated by the stimuli.

The values for σ_j and σ_{jk} were calculated as indicated in [5], that is:

$$\sigma_j = \frac{n}{V_j \sum \left(\frac{1}{V_j} \right)} \quad (24)$$

where the summation sign refers to all the j values, and V_j stands for the standard deviation of the X_{ij} values, that is: $V_j = \sqrt{\left[\frac{\Sigma X_{ij}^2}{m-1} - \left(\frac{\Sigma X_{ij}}{m-1} \right)^2 \right]}$. In this equation the summation sign applies to all the i values. S_j and S_{jk} were determined from the experimental values using eqn. (11) and (13). S_{jk}^* may be calculated by using eqn. (1) or determining first the values of $X_{ijk}^* \sigma_{jk}^*$ as indicated in (19) and then applying formula (13) with the pertinent changes. The value of σ_{jk}^* was calculated by using formula (20). r_{jk} was estimated by correlating the original scores. For this purpose the successive intervals were assigned arbitrary values of

Prediction of Scale Values for Combined Stimuli

TABLE I. GRAPHIC RATING SCALE

Decimal point omitted

		1	2	3	4	5	6	7	8	9
Napoleon	A*	035	024	047	029	159	170	129	123	282
Hitler	T*	047	027	030	045	174	137	135	144	261
Roosevelt	A	018	012	035	018	212	194	194	176	141
Mussolini	T	036	031	041	066	263	192	156	121	094
Napoleon	A	000	006	006	029	176	282	194	176	129
Sarmiento	T	008	011	019	038	219	213	197	170	125
Hitler	A	018	012	035	053	212	253	206	123	088
San Martin	T	042	033	040	064	250	177	152	127	115
Bolivar	A	000	024	035	094	294	229	176	094	053
Mussolini	T	046	041	053	084	312	193	137	088	046
Hitler	A	006	000	018	024	165	176	170	165	276
Roosevelt	T	028	023	030	048	205	170	159	153	184
Sarmiento	A	024	012	047	059	247	229	206	106	071
San Martin	T	054	040	046	075	265	179	143	111	087
Napoleon	A	000	006	006	006	094	182	176	153	376
Roosevelt	T	010	010	015	028	152	152	173	186	274
Sarmiento	A	012	035	029	076	300	247	176	088	035
Mussolini	T	060	049	062	097	323	185	120	071	033
San Martin	A	024	024	024	059	223	212	182	159	094
Bolivar	T	035	031	038	065	260	189	158	125	099
Napoleon	A	018	006	029	047	153	206	194	153	194
Mussolini	T	058	034	039	059	215	155	138	133	169
Roosevelt	A	000	006	006	024	135	229	265	188	147
Sarmiento	T	019	020	031	051	242	201	175	146	115
Napoleon	A	000	000	018	006	135	218	194	241	188
San Martin	T	010	012	020	036	200	190	191	175	166
Bolivar	A	024	035	053	029	247	170	194	188	059
Sarmiento	T	062	038	050	068	250	169	139	115	109
Roosevelt	A	000	000	000	006	094	265	247	176	212
San Martin	T	007	011	020	040	234	220	200	162	106
San Martin	A	029	018	047	053	300	247	165	082	059
Mussolini	T	046	041	055	091	323	196	131	079	038
Roosevelt	A	000	000	006	024	135	294	235	135	170
Bolivar	T	011	013	022	041	222	199	194	164	134
Hitler	A	012	018	053	059	218	212	194	153	082
Sarmiento	T	051	036	047	069	257	177	145	118	100
Bolivar	A	006	000	000	006	147	288	153	188	212
Napoleon	T	010	012	018	035	193	184	188	178	182
Hitler	A	000	035	029	094	147	223	200	165	106
Bolivar	T	036	030	038	060	245	186	157	129	119
Hitler	A	112	018	035	065	129	129	159	153	200
Mussolini	T	145	050	053	068	200	128	108	099	149
Mean Discrepancies (per column)		024	014	010	022	050	040	039	027	036

Total number of cells = 189. Mean discrepancy = 029 (total). A* = actual proportions; T* = theoretical proportions.

TABLE II. METHOD SUCCESSIVE INTERVALS

Decimal point omitted

		1	2	3	4	5	6	7	8	9
Cervantes	A*	024	024	029	082	129	288	188	182	053
Hitler	T*	028	026	040	070	145	183	174	235	099
Napoleon	A	000	000	006	006	029	141	253	306	259
Roosevelt	T	009	010	017	033	083	132	156	305	255
Roosevelt	A	000	006	024	012	059	194	270	341	094
Bolivar	T	005	009	020	042	124	201	217	293	089
San Martin	A	012	018	029	112	129	329	247	106	018
Mussolini	T	020	030	054	099	218	245	178	139	017
Napoleon	A	006	000	006	035	094	223	312	212	112
Sarmiento	T	010	013	025	049	127	189	197	282	108
Hitler	A	006	035	065	070	094	188	265	206	070
Bolivar	T	033	029	044	073	144	185	169	226	097
Sarmiento	A	024	035	059	118	170	294	188	100	012
Mussolini	T	038	041	066	107	204	221	159	141	023
Roosevelt	A	006	000	000	024	065	241	235	318	112
San Martin	T	007	011	021	048	131	203	212	280	087
Hitler	A	006	006	006	012	053	123	247	306	241
Sarmiento	T	048	036	050	078	151	181	158	206	092
San Martin	A	029	035	047	094	083	259	235	182	035
Bolivar	T	026	029	045	082	170	212	181	201	054
Napoleon	A	012	006	041	076	076	182	247	276	083
Mussolini	T	038	029	041	063	131	158	158	237	145
Hitler	A	006	035	047	059	106	194	212	194	147
Roosevelt	T	035	024	035	055	112	144	147	248	200
Roosevelt	A	018	041	083	070	229	229	200	106	024
Mussolini	T	026	026	040	069	144	187	174	237	097
Roosevelt	A	006	047	029	070	229	282	212	112	012
Sarmiento	T	020	021	034	061	135	185	181	254	109
Bolivar	A	000	006	024	041	135	206	294	182	112
Mussolini	T	023	031	052	097	203	238	177	156	023
Napoleon	A	000	029	076	094	176	323	167	100	035
San Martin	T	009	012	024	045	119	181	193	292	125
Bolivar	A	000	006	000	047	088	259	194	259	147
Sarmiento	T	034	033	052	087	176	205	171	190	052
Sarmiento	A	012	035	053	083	147	265	200	159	047
San Martin	T	047	038	051	091	168	194	163	184	058
Hitler	A	024	018	053	094	123	229	253	159	047
Mussolini	T	127	049	054	075	120	131	114	176	154
Bolivar	A	123	035	035	041	083	165	165	194	159
Napoleon	T	005	009	018	040	112	183	204	310	119
Hitler	A	000	012	018	029	070	235	223	265	147
San Martin	T	036	031	044	076	150	183	168	220	092
Napoleon	A	018	029	024	059	088	206	282	253	041
Hitler	T	040	026	033	053	103	127	138	241	239
Mean Discrepancies per)										
column)		025	014	017	021	054	057	063	058	056

Total Number of cells = 198. Mean discrepancy = 041 (total). A* = Actual proportions; T* = Theoretical proportions.

Prediction of Scale Values for Combined Stimuli

TABLE III. MULTIPLE CATEGORY METHOD

Decimal point omitted

		1	2	3	4	5
Roosevelt	A*	012	041	076	429	441
San Martin	T*	034	085	129	399	353
Mussolini	A	024	194	206	424	153
Sarmiento	T	094	180	210	390	126
Roosevelt	A	029	112	141	406	312
Mussolini	T	068	117	148	375	292
Bolivar	A	059	094	153	394	300
Hitler	T	101	127	142	338	292
Napoleon	A	018	041	118	388	435
Sarmiento	T	022	075	136	448	319
Mussolini	A	035	206	165	441	153
San Martin	T	096	164	192	385	163
Bolivar	A	059	147	147	418	229
Sarmiento	T	071	145	185	409	190
Roosevelt	A	006	059	035	247	653
Napoleon	T	026	057	089	322	506
Hitler	A	041	141	106	418	294
San Martin	T	108	125	138	325	304
Roosevelt	A	012	076	106	382	424
Bolivar	T	029	082	131	416	342
Mussolini	A	182	100	059	294	365
Hitler	T	211	126	116	244	303
Napoleon	A	000	059	147	394	400
Bolivar	T	023	073	126	424	354
Sarmiento	A	059	129	165	429	218
Hitler	T	106	133	149	342	270
San Martin	A	065	141	206	370	218
Sarmiento	T	067	142	185	414	192
Napoleon	A	047	100	118	365	370
Mussolini	T	081	115	137	346	321
San Martin	A	012	070	147	365	406
Napoleon	T	027	075	125	408	365
Roosevelt	A	029	070	088	306	506
Hitler	T	082	095	111	296	416
Bolivar	A	047	147	135	335	335
San Martin	T	066	133	173	407	221
Napoleon	A	041	082	082	288	506
Hitler	T	086	094	106	283	431
Roosevelt	A	018	059	129	447	347
Sarmiento	T	038	093	139	408	322
Bolivar	A	035	182	200	424	159
Mussolini	T	092	168	197	390	153
Mean Discrepancies (per column)		032	018	024	043	055

Total number of cells = 105. Mean discrepancy = 034 (total). A* = actual proportions; T* = theoretical proportions.

1,2,3,..., m . For each subject and for each stimulus a value corresponding to the location of the mark indicating his choice was thus obtained and from these values the correlations were calculated.¹

TABLE IV.

Stimuli	GRS		MSI		MCM	
	S_j	σ_j	S_j	σ_j	S_j	σ_j
Roosevelt	.417	.877	.521	.994	.511	.980
Hitler	-.013	1.222	.052	1.408	.018	1.501
Mussolini	-.627	1.061	-.526	1.026	-.532	1.033
Napoleon	.685	1.087	.706	1.160	.587	1.039
San Martin	-.124	.771	-.212	.797	-.136	.766
Sarmiento	-.294	1.032	-.338	1.034	-.284	.835
Bolivar	-.043	.951	-.138	.812	-.163	.847
Cervantes	...	...	-.064	.769	...	...
Σ	.001	7.001	.001	8.000	.001	7.001

V. EXPERIMENTAL RESULTS

Table IV gives the S_j and σ_j values for GRS, MSI and MCM. The upper part of the cells in Tables I, II, III give the p_{ijk} values for GRS, MSI and MCM. Table V gives the S_{jk} and σ_{jk} values for GRS, MSI and MCM, and Table VI gives the S_{jk}^* and σ_{jk}^* values. Plotting eqn. (16) fig. 1, 2 and 3 were obtained. The correlations between S_{jk} and S_{jk}^* are: 0.93 for GRS, 0.97 for MSI and 0.96 for MCM. Plotting eqn. (8) for all the stimuli, parallel straight lines were obtained. It can be demonstrated that

$$e = \sqrt{\frac{n^2 \sum \left(\frac{1}{V_{jk}} \right)^2}{n'^2 \sum \left(\frac{1}{V_j} \right)^2}} \quad (25)$$

where V_j represents the standard deviation of the X_{ij} values and V_{jk} the standard deviation of the X_{ijk} values. In eqn. (25) the summation sign in the numerator and denominator refer respectively to the jk and j values. By plotting σ_{jk}^* against σ_{jk} Fig. 4, 5 and 6 were obtained. The correlations between σ_{jk}^* and σ_{jk} are 0.95 for GRS, 0.92 for MSI and 0.92 for MCM. The values of r_{jk} are presented in Table VII.

Using formula (23) the X_{ijk}^* values were estimated, and the corresponding p_{ijk}^* were calculated from a probability integral table. The values of these proportions are given in the lower part of the cells in Tables I, II and III. The average discrepancies between p_{ijk} and p_{ijk}^* are as follows: .03 for GRS, .04 for MSI, and .03 for MCM, with the corresponding standard deviations, .04, .06 and .04, respectively.

¹ The correlation also may be determined if after computation of L_i , the mid values of the intervals on the psychological continuum are used instead of the arbitrary numbers previously described. L. V. Jones, working independently of the author, used both procedures and did not find any major difference between them [4].

TABLE V.

	GRS		MSI		MCM	
	S_{jk}	σ_{jk}	S_{jk}	σ_{jk}	S_{jk}	σ_{jk}
Roosevelt San Martin	.362	.901	.301	.885	.315	.906
Mussolini Sarmiento	-.451	.840	-.560	.633	-.432	.737
Roosevelt Mussolini	-.003	1.044	.008	.988	-.080	.923
Bolivar Hitler	-.134	1.013	-.150	1.037	-.120	1.035
Napoleon Sarmiento	.091	.833	.199	.841	.263	.948
Mussolini San Martin	-.434	.926	-.380	.790	-.446	.785
Bolivar Sarmiento	-.303	1.005	-.300	.913	-.303	.956
Roosevelt Napoleon	.667	1.046	.724	.919	.873	1.227
Hitler San Martin	-.210	.947	-.135	.949	-.137	.974
Roosevelt Bolivar	.172	.890	.264	.953	.181	.903
Mussolini Hitler	-.118	1.622	-.150	1.614	-.110	1.742
Napoleon Bolivar	.304	.824	.282	.972	.154	.855
Sarmiento Hitler	-.224	.943	-.330	.973	-.309	.932
San Martin Sarmiento	-.324	.943	-.284	.952	-.364	.949
Napoleon Mussolini	.113	1.092	.043	1.149	.040	1.093
San Martin Napoleon	.314	.951	.255	.988	.117	.874
Roosevelt Hitler	.365	1.064	.476	1.230	.411	1.170
Bolivar San Martin	-.206	1.019	-.151	.918	-.111	1.055
Napoleon Hitler	.194	1.393	.455	1.343	.413	1.281
Roosevelt Sarmiento	.190	.863	.090	.932	.076	.872
Bolivar Mussolini	-.366	.841	-.417	.855	-.430	.786
Cervantes Hitler	...	...	-.240	.964	...	...

TABLE VI.

	GRS		MSI		MCM	
	S_{jk}^*	σ_{jk}^*	S_{jk}^*	σ_{jk}^*	S_{jk}^*	σ_{jk}^*
Roosevelt San Martin	·146	·812	·146	·875	·187	·921
Mussolini Sarmiento	—·461	·882	—·440	·896	—·408	·821
Roosevelt Mussolini	—·105	·960	—·011	1·039	—·011	·994
Bolivar Hitler	—·028	1·006	—·052	1·073	—·073	1·108
Napoleon Sarmiento	·196	·841	·176	·939	·152	·811
Mussolini San Martin	—·376	·864	—·378	·805	—·334	·883
Bolivar Sarmiento	—·168	1·081	—·246	·975	—·224	·862
Roosevelt Napoleon	·551	1·019	·605	1·113	·549	1·045
Hitler San Martin	—·069	1·025	—·088	1·073	—·059	1·154
Roosevelt Bolivar	·187	·880	·183	·860	·174	·880
Mussolini Hitler	—·320	1·433	—·246	1·526	—·257	1·531
Napoleon Bolivar	·321	·927	·276	·902	·212	·853
Sarmiento Hitler	—·154	1·028	—·151	1·121	—·132	1·085
San Martin Sarmiento	—·209	1·010	—·284	1·034	—·210	·853
Napoleon Mussolini	·029	1·186	·082	1·186	·028	1·083
San Martin Napoleon	·280	·910	·239	·958	·226	·889
Roosevelt Hitler	·202	1·068	·278	1·270	·265	1·258
Bolivar San Martin	—·084	·966	—·184	·947	—·150	·890
Napoleon Hitler	·336	1·298	·370	1·369	·303	1·314
Roosevelt Sarmiento	·062	·916	·083	1·024	·114	·904
Bolivar Mussolini	—·335	·890	—·340	·844	—·348	·859
Cervantes Hitler	...	...	—·015	1·048	...	...

Prediction of Scale Values for Combined Stimuli

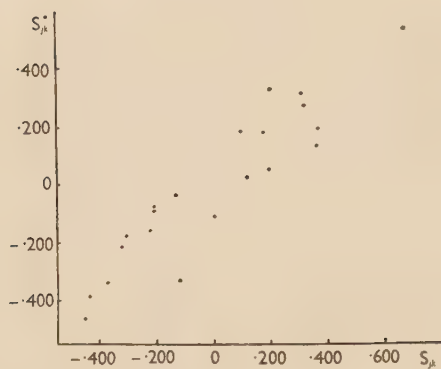


Figure 1. GRS

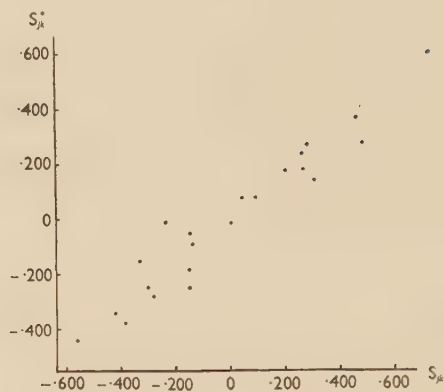


Figure 2. MSI

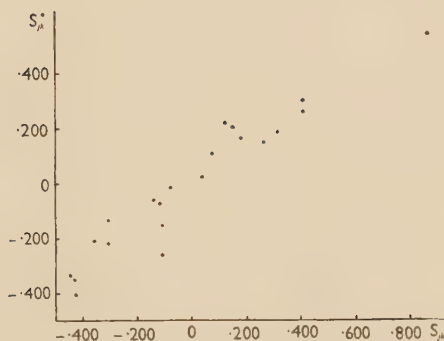


Figure 3. MCM

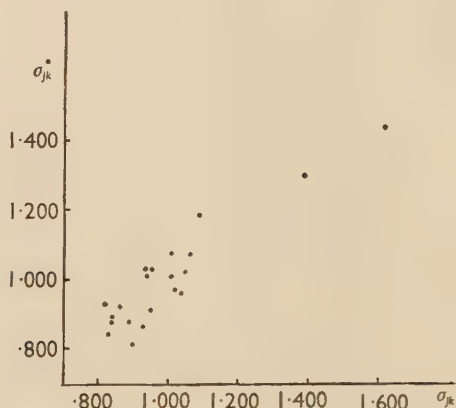


Figure 4. GRS

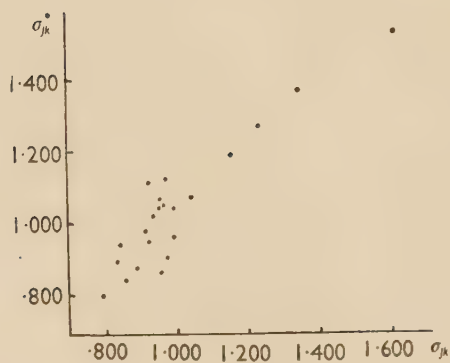


Figure 5. MSI

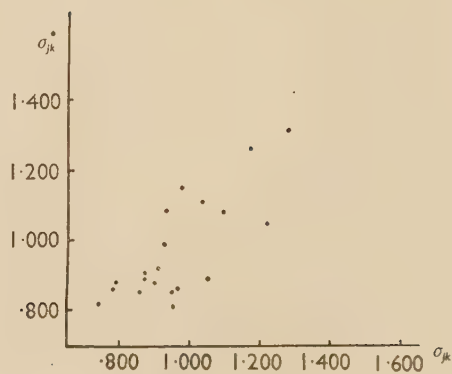


Figure 6. MCM

VI. DISCUSSION

The previous experimental results indicate that our data show a linear relationship between the scale value of the single stimuli and the value of their combinations when given in groups of two. Tables I, II, III show the accuracy of the prediction. The value of the average discrepancies compare favourably with those presented in the usual psychophysical studies, where the question of predicting from single to double stimuli is not involved. It would be interesting to know whether the kind of relationship here described also holds when the single stimuli are combined into larger groups or when other types of stimuli are used.

TABLE VII.

		Roosevelt	Hitler	Mussolini	Napoleon	San Martin	Sarmiento	Bolivar	Cervantes
Roosevelt		—							
		—							
		—							
Hitler	MCM	·170	—						
	GRS	·123	—						
	MSI	·157	—						
Mussolini	MCM	·145	·703	—					
	GRS	·075	·741	—					
	MSI	·118	·738	—					
Napoleon	MCM	·255	·229	·282	—				
	GRS	·181	·396	·347	—				
	MSI	·124	·197	·242	—				
San Martin	MCM	·295	·123	·112	·119	—			
	GRS	·070	·125	—·042	·033	—			
	MSI	—·001	—·081	—·193	—·021	—			
Sarmiento	MCM	·159	—·075	—·105	—·135	·332	—		
	GRS	·011	—·088	—·215	—·305	·375	—		
	MSI	·079	—·134	—·200	—·229	·342	—		
Bolivar	MCM	·085	—·035	—·032	—·052	·427	·232	—	
	GRS	·024	—·069	—·137	—·089	·385	·313	—	
	MSI	—·051	—·089	—·123	—·150	·465	·170	—	
Cervantes	MSI	—·005	—·115	—·043	—·025	—·060	·015	·045	—

The scale values for such items as food, clothing, etc., where price or other considerations may partially determine the choice and establish a ceiling, may or may not follow the type of relations described in this paper. The problem should be investigated since it may have considerable theoretical and practical implications. It would scarcely be possible to enumerate all the situations in which the method here indicated may be used. It may have application in studies of public opinion, for instance prediction of elections (when several candidates are involved), in market research, in studies of decision making when several elements have to be taken into consideration, in determining changes in degree of affect under varied circumstances, and so forth. It may prove a labour-saving device in those cases where it is difficult or costly to present two stimuli simultaneously. Further, it provides a way of estimating the psychological magnitude of different parts on the continuum, and therefore of establishing the equivalence of different segments on the psychological continuum.

It is interesting to notice that the scale values for the single stimuli have substantial correlations among themselves and that these correlations cover a considerable span of values. The assumptions are usually made that the correlations between stimulus judgements are all equal or zero, and the values obtained for single stimuli by using the paired comparison method rest upon these assumptions. The credibility of these values is therefore implicated in the validity of the assumptions. Correlations between stimulus judgements *qua* responses have been demonstrated in psychophysical experiments [1] and this phenomenon may be related to the correlations we have obtained.

REFERENCES

1. Goldiamond, I. (1955). *Serial effects as a function of the type of indicator used.* (Abstract submitted to Midwestern Psychological Association.) (Available from American Documentation Institute.)
2. Guilford, J. P. (1931). 'The prediction of affective values', *Amer. J. Psychol.*, XLIII, 469-478.
3. Guilford, J. P. (1933), Spence, W. 'The affective value of combinations of odors.' *Amer. J. Psychol.*, XLV, 495-501.
4. Jones, L. V. (1954). 'Correlation coefficients derived from raw and scaled rating data.' *Psychometric Laboratory*, Univ. of Chicago (mimeo.).
5. Rimoldi, H. J. A. and Hormaeche, M. (1955). 'The law of comparative judgement in the successive intervals and graphic rating scale methods.' *Psychometrika*, XX, 317-318.

WEIGHTING RESPONSES TO ITEMS IN ATTITUDE SCALES

By PATRICK SLATER

Institute of Psychiatry, Maudsley Hospital, University of London
(Formerly Science 4, Air Ministry)

Abstract. A metrical treatment of the problem of weighting responses to items in an attitude scale is described. The procedure proposed is developed from a method suggested by Guttman in 1941 and since discarded. A test for scalability is introduced. Procedures are described for carrying out the requisite iterations either with pencil and paper or with the Hollerith machine. Various methodological problems are also discussed.

I. The Psychological Context. II. Statistical Problems. III. General Treatment of Data. IV. Computation of Preliminary Weights. V. The Iterative Process Proposed: (a) with Pencil and Paper; (b) with the Hollerith Machine. VI. Statistical Devices and Inferences. VII. Discussion. VIII. Criteria of Scalability. IX. The Choice of Weights.

I. THE PSYCHOLOGICAL CONTEXT

A technique for measuring morale is urgently needed by such bodies as the Air Ministry because the administrative changes they have to consider are often intended to improve morale. To find out whether the effects intended are in fact obtained, measurements must be taken before and after the change, and results compared.

In constructing scales for such a purpose the relevant branch of the Ministry (Science 4) has benefited greatly from experience in the United States and Canada. There now appears to be a general agreement that morale should be regarded, not as a unitary trait, but as a composite of several attitudes which require to be separately assessed. The main questionnaire now used in the R.A.F. is composed of six scales: the most important are designed to assess (i) the airman's valuation of his own rôle as compared with that of his contemporaries in civil life, (ii) his identification with the station where he is serving, (iii) his valuation of it in comparison with other stations, and (iv) his identification with the R.A.F. Items belonging to the same scale often seem similar; but when mixed with others from different scales they provide a sufficient variety to give the airman an interesting and easy task. The time allowed is enough for everyone to choose a response to each item. The material and procedure are illustrated in Table I. For purposes of formal discussion the 'attitude' measured by a particular scale may be defined as the degree of favourableness shown by the testee towards some general object with which the scale is concerned.

II. STATISTICAL PROBLEMS

The experimenter who desires to construct a scale to measure some particular attitude will begin by making a psychological analysis of the attitude. Men in the groups concerned are interviewed informally to discover what they feel and how they express themselves. A preliminary selection of items can then be made: those most closely related to the defined attitude are selected for inclusion in the scale, and the rest relegated to other scales or discarded. The alternative responses to be offered are chosen from the same material, and ranked in a preliminary order of favourableness (not necessarily the order in which they will be presented). The draft scale can then be tried with a representative sample.

Weighting Responses to Items in Attitude Scales

The frequencies of the responses are then used to answer the following questions: (a) Is the attitude scalable as a whole? (b) Is each particular item suitable for inclusion? (c) Are the alternative responses to the same item distinct from one another?

The prior doubt whether the items relate to the same attitude may arise in more than one way. In an ill-constructed scale items relating to several distinct attitudes may be mixed or a completely inconsistent assortment may be assembled. But even a scale known to be reliable may fail if applied to a group where the attitude has not developed. Thus, when the R.A.F. scale for comparing airmen's with civilian status was applied to National Service men, inconclusive results were obtained, because everyone in the age group is affected in some way by the call-up.

TABLE I. ILLUSTRATIVE ITEMS AND RESPONSES FROM AN ATTITUDE SCALE

Instructions: Read the questions and all the alternative answers. Put a tick by the answer that most nearly represents your own opinion.

I. *When he leaves this station, an entrant*

- A. Has a big advantage over other people in the service;
- B. Has some advantage over other people in the service;
- C. Is no different from anybody else;
- D. Is at something of a disadvantage in comparison with other people in the service;
- E. Is at a big disadvantage.

II. *If you were posted to another station and heard that another entrant from this station whom you did not know had also been posted there, would you*

- A. Certainly try to look him up;
- B. Probably try to look him up;
- C. Not care whether you came across him or not;
- D. Probably not try to look him up;
- E. Certainly not try to look him up?

The two items in Table I will illustrate problems of the second and third kind. It is open to doubt whether both relate to the same attitude: for, even if a man has a low opinion of his station, he may have some fellow-feeling for others who have been there and want to 'look them up'. It is also doubtful whether the order of favourableness of the responses to item II is correct. It is obviously quite conceivable that response C might prove less favourable than D, and approximately equivalent to E.

If the attitude is scalable as a whole, problems concerning particular items and particular responses may both be reducible to that of finding the optimum weight for each response. Thus, if the whole item is unrelated to the attitude, a procedure giving the optimum weights will assign equal weight to all its responses, and so exclude it from the scale, since its contribution to the total variance will be zero. If the order originally assigned to the responses is wrong, the optimum weighting will correct it. A test for scalability, and a procedure for computing optimum weights, should suffice for treating the chief problems that arise.

III. GENERAL TREATMENT OF DATA

The approach here adopted is similar to that described by Guttman [5] and explained in the notes on Table II.

If an item is presented to a group of men whose attitudes have been measured already, the degree of favourableness indicated by a particular response is best estimated, according to the method of least squares, as the average score of the men choosing it. This leads to equations of type 2 for estimating weights for responses (w). Conversely, if the degrees of favourableness indicated by the responses are already known, the best estimate of a man's attitude is obtained by averaging the weights of the responses he chooses. This leads to equations of type 1 for estimating scores (z).

TABLE II. RESPONSES CHOSEN BY u INDIVIDUALS FROM A SCALE CONTAINING m ITEMS AND n RESPONSESTHE MATRIX M

Scale		Individual (i)					
Item	Response (j)						
		(1)	(2)	(3)	(4)	...	u
I.	A	...	1	...	...	...	...
	B	1	...	...	1	...	...
	C	...	...	1	...	...	...
II.	...	...	...	...	...	...	...
	A	...	1	...	1	...	...
	B	...	...	1	...	...	...
	C	1	...	...	...	...	...
	...	...	...	...	...	...	...

Explanation. The response to an item chosen by an individual is recorded by a unit under his code number opposite his response: e.g., the entries in column (1) show that individual 1 has chosen the B response to item I, C to item II, etc. With m items in the scale, there are m units in every column. Numbers in rows vary; N_j denotes the number in row j .

Scores (z) for the individuals may be obtained by summation if weights (w) for the responses are known. The w may be written as a row vector and multiplied into M giving

$$z = wM \quad (1)$$

z is then a row vector, with transpose z' .

Conversely if z is known, weights for the responses may be obtained by averaging. Writing the reciprocals of the values of N_j as a diagonal matrix, D^{-1} , we obtain

$$w' = D^{-1}Mz'. \quad (2)$$

Usually z and w are both unknown; but preliminary estimates of w are deducible from hypothesis. Entering them in eqn. (1) starts an iterative process leading to successive estimates of z and w .

Guttman (1941) argues that optimum values of z and w are reached when the conditions $z = wM$ and $p_w = D^{-1}Mz'$ are satisfied simultaneously (p being a constant of proportionality), i.e., when the iterative process stabilizes. At this point the variances between persons and between responses are both maximized relatively to the total variance of the weighted responses. He does not argue that there is a unique set of optimum values.

Independent measurements of z are generally unobtainable, but preliminary estimates of w can be made. Introducing the weights (w) into equation (1) starts an iterative process which leads to first estimates of z , then to revised estimates of w , and so on. Guttman argues that it terminates in a set of optimum weights for both, connected by a simple relation.

The possibility of finding more than one such set for a single scale cannot be excluded: in theory, it is possible for the same scale, differently weighted, to measure different attitudes. But the psychological context usually makes such a procedure not worth considering. To measure a second attitude, a second scale is constructed. The hypothesis that the two are not independent may be tested by carrying out critical experiments on samples with varying composition, and observing their covariance. Two or more scales, each with a narrow definition or range, may then be combined into a broader questionnaire.

IV. COMPUTATION OF PRELIMINARY WEIGHTS

The relative weights for the alternative responses to the same item are deduced from the psychological hypothesis, which ranks them in order of favourableness, and from their relative frequency of occurrence in the sample. This permits us to calculate the normal deviate values. The rough method given by Garrett [3] is convenient, though for accuracy Karl Pearson's original procedure [7] is to be preferred.

Weighting Responses to Items in Attitude Scales

Either method gives for each item a set of preliminary weights (w_0) having a mean of 0 and a standard deviation of approximately 1. They take none of the observed relations between the items into account; but, if the psychological hypothesis is correct, when the items are so weighted, they should tend to correlate positively. This can be tested without calculating correlations. The necessary data are provided by the distribution of the first estimates of scores (z_1) extracted by iteration. Their total variance (V_{z_1}) = the sum of the variances of the items (ΣV_{w_0}) + twice the sum of their covariances (ΣW_{w_0}), where V denotes the variance of the quantity designated by the lower case letter and W (two V 's) denotes the covariance. ΣV_{w_0} is thus the expected value of V_{z_1} on the null hypothesis. If there is a general tendency to positive correlation, V_{z_1} must exceed it significantly. This can be tested by entering the usual table of variance ratios with $V_{z_1}/\Sigma V_{w_0}$ as a variance ratio with $u-1$ d.f., or preferably $(u-1)V_{z_1}/\Sigma V_{w_0}$ as a chi-squared. When the result is satisfactory, iteration is allowed to proceed.

V. THE ITERATIVE PROCESS PROPOSED

Methods have been developed for performing the requisite iterations both by pencil and paper and by the Hollerith machine.

(a) *With Pencil and Paper.* One sheet (the u -sheet) will be needed for each person and a larger one (the m -sheet) for each item. They correspond with the columns and the rows of the matrix M in Table II. In the first column of a u -sheet the individual's responses are listed. The list for the first person in Table II would begin

Item	Response
I	B
II	C

and continue with the rest of the information in the first column of the Table.

The m -sheet will be larger, since it must be divided vertically into blocks for each alternative response. The first column of entries in a block is the list of code numbers for the individuals choosing that response. Thus the sheet for item I in Table II would start with entries of 2 in block A, 1 and 4 in block B, etc., thus:

		Item I			
		A	B	C	...
Persons	2	1	3	...	...
	...	4	...	...	...

and continue with all the information in the rows of Table II relating to item 1.

Counting the entries in the blocks will give the frequency distributions needed to calculate preliminary estimates for the weights of the responses. They are entered on the u -sheets: thus, the weights for I B, II C, etc., would form the second column of entries on the sheet for individual 1. And their sum will be the first estimate of his score.

After the distribution of the scores has been extracted for the test of scalability, they are entered on the m -sheets, e.g., individual 1's score would appear opposite 1 under block B in the sheet for item I, and similarly on the other sheets. The average of the scores in a block provides the next estimate of the weight for that response, ready for transfer back to the u -sheets, etc.

(b) *With the Hollerith Machine.* The Hollerith will require 3 packs of cards:

- A pack of cards, $m \times u$ in number, one for each response, i.e., for each unit entry in Table II;
- A second pack of cards, u in number, one for each person, i.e., for each column in Table II;
- A third pack, n in number, one for each alternative response, i.e., for each row in Table II.

The B pack is punched first by hand: if the lay-out of the questionnaire is suitable, this can be done directly without prior coding. The entries include the code number (and other data if desired) for the individual, and the code numbers of the responses he has chosen, one column being allocated to each item.

The A pack may be punched from the B pack by machine. Its entries are only the code numbers of the individual and the responses concerned. A sorting of items by responses provides the frequency distributions for the preliminary weights. While they are calculated, the Hollerith process must be suspended. They are then punched by hand on to the C pack and transferred to the A pack by machine. This can then be sorted according to individuals, accumulating the weights of each man's responses and entering their total on his B card as the first estimate of his score.

Here the test for scalability intervenes. If its outcome is not in doubt, the Hollerith process may be allowed to run on. The scores are then machine-punched on to the A cards, and accumulated by response to provide the data required for revising the weights. After final scores have been entered on to the B pack, it will be available for the subsequent stages of the investigation.

VI. STATISTICAL DEVICES AND INFERENCES

Both the pencil and paper and the Hollerith processes run most smoothly with weights in the range from 0 to 100; and weights calculated by Pearson's or Garrett's method can be adjusted to this requirement. Their relative variances remain unaffected if they are all multiplied by a constant, p , or assigned different origins by adding a set of constants, $k_1 \dots k_m$, one for each item. To obtain the desired range convenient values of p usually lie between 15 and 20; the k 's are chosen so that the least favourable response to each item receives a weight of 0. The observed values of Vw_0 should be computed after these adjustments, partly as a check and partly in order to obtain an accurate value of ΣVw_0 for the test of scalability.

A similar adjustment will be found convenient after the revised weights have been obtained from eqn. (2) at the end of the first cycle of iteration. The changes occurring during iteration are most clearly observable if ΣVw is scaled to the same quantity at the start of each cycle. To scale ΣVw_1 down to ΣVw_0 , p can be calculated as $\sqrt{\{\Sigma Vw_0 / \Sigma Vw_1\}}$.

TABLE III. PROPORTIONAL CHANGES IN THE VARIANCE OF THE WEIGHTS OF DIFFERENT ITEMS IN A DRAFT SCALE

Item	Proportional Variances			
	w_0	w_1	w_2	w_3
I	·9808	·3422	·2177	·1680
II	1·0044	1·2963	1·3656	1·3433
III	·9992	1·0385	1·0698	1·0761
IV	1·1175	·7743	·7681	·7331
V	·8725	1·0679	1·1589	1·1586

Table III illustrates the effects of the process. Here the successive values of ΣVw have all been scaled down to m . Item I is obviously unsatisfactory: the weights for its alternative responses converge, and its variance diminishes rapidly. Item II follows the opposite course. Such evidence will usually suffice to support conclusions about the relative merits of the items, and to provide reasonable weights for the alternative responses, though further statistical refinements may be desirable. In particular, it would be an advantage to be able to test whether the correlation between an item and the whole scale is significant, when optimum weights are used.

Weighting Responses to Items in Attitude Scales

Failure to Stabilize. The iterative process does not always converge on an optimum set of weights. In one of our own investigations nearly all the items in the scale concerned evoked predominantly favourable responses from the sample tested. The skewness was passed on to the z scores, and back again to the successive estimates of w . The w 's showed no tendency to stabilize; and the distribution of z became increasingly asymmetrical at an accelerating rate as iteration proceeded. Nevertheless, even under these unfavourable conditions, it was found that the proportional variance of the items stabilized rapidly, like those in Table III. Practical conclusions were not precluded, though the weights finally allotted to the items could scarcely be dignified by the title 'optimum'.

VII. DISCUSSION

Comparison with Professor Guttman's Procedure. In his later treatment of problems of this kind Professor Guttman discarded the methods he described in 1941 in favour of others more highly specialized. It may seem odd to revert to his earlier methods, but there seem good practical reasons for so doing. Guttman's more recent methods [8] require a pair of scalogram boards, containing long narrow trays to correspond with the rows and columns of M (our Table II). By this means responses and individuals can both be ranked in order of favourableness. When this operation is completely successful, the rank in which a person is placed would indicate precisely which responses he chose. Thus 100 per cent. reproducibility would be achieved. In practice this ideal is not attained; but scales giving 95 per cent. reproducibility are frequently reported, though less than 90 per cent. reproducibility is viewed with disfavour. The iterative process employed is not completely automatic, but provides scope for intuitive judgements. Alternative responses to the same question may be combined during iteration to improve reproducibility; this process is irreversible. A pencil-and-paper method has also been described.

The most radical change is from metrical terms to terms of rank order. Guttman's reason for this have not, so far as I am aware, been fully stated. Methodologically, attitudes do not seem to differ from other psychological traits that are usually treated as continuous variables. Suppose a man contented with his service conditions were to be severely punished for what he considered a trivial offence. A change in his attitude might result, perhaps a complex one: his favourableness towards the officer reporting him might decline faster than his favourableness towards the service as a whole. Metrical terms seem thoroughly appropriate for describing such phenomena. In any case it seems doubtful whether such terms can be systematically avoided. If the conclusion is reached that an attitude is scalable, weights will be assigned to the responses and scores calculated for persons. Guttman fully agrees that this is the final result to be reached. Does it not imply that a proper question to raise is what are the optimum weights for the responses—a question which in essence is numerical?

In other psychological inquiries metrical methods, when possible, are preferred to methods of rank order. With the latter the main difficulty lies in equating differences in rank. Few of the advanced methods of statistical analysis are directly applicable. However, several procedures for converting rank orders into measurements have been described and some are widely used [2]. A case for reversing this preference needs to be strong.

VIII. CRITERIA OF SCALABILITY

If it is allowed that each item defines, however approximately, a dimension of variance, the geometric picture presented by all the items in a scale is the familiar one of variation occurring in an m -dimensional space. Or if the alternative responses to an item do not necessarily belong in a single continuum, the variation may extend into $n-m$ dimensions. Similarities with the picture presented by factor analysis are noteworthy: this has been pointed out by Burt [1], who draws attention to the close similarity between 'scale analysis' and certain of its formulæ and factor analysis by 'weighted summation'. In the analysis of an attitude scale the hypothesis isolates variance in a single dimension for consideration, namely, that from the all-positive to the all-negative hyper-quadrant in the m -space, affirming that it is significant. The null hypothesis can be used to test it.

I can quote no reference for the proposed test of scalability; but I am indebted to Professor M. S. Bartlett for advice that it is appropriate. The essential conditions are that the preliminary weights (as defined by the hypothesis) should be used, and that they should provide approximately equal variances for the items. The latter condition is acceptable psychologically, since as a rule no prior assumptions about the relative importance of the items are introduced. To be exact, the ratio $(u-1)V_z/\Sigma Vw_0$ differs from chi-squared in having an upper limit at $m(u-1)$ whereas chi-squared has none; but with the values of m and u usually involved this is negligible.

All methods of estimating the reliability of a test without repeating it are based on the observed relations between its variance and the variance of its component items; i.e. they are derived from the same terms as the ratio $V_z/\Sigma Vw$. Estimates in general use have the advantage of satisfying the psychologist's regular preference for correlation coefficients, especially when they are positive and substantial. But their errors are unknown. The plain ratio is easiest to derive from the data, and leads directly to an efficient test of significance. On balance it seems worth recommending. But a Kuder-Richardson reliability coefficient, case II [6], or even a split-half reliability could also be derived.

The coefficient of reproducibility used by Guttman appears to be only indirectly and vaguely related to the question whether the preliminary weights deduced from the psychological hypothesis justify combining the items into a scale. It is primarily a descriptive measurement for the scale's reliability after an approximation for the optimum weights has been reached.

IX. THE CHOICE OF WEIGHTS

Once scalability is demonstrated, the next question is to find the optimum weights. Here a difference of opinion may arise over the definition of 'optimum'. My own view is that the optimum weights should be those which maximize the relative variance in the dimension defined by the psychological hypothesis, and that this is equivalent to maximizing the ratio $V_z/\Sigma Vw$ for the given m . If so, this would agree with Guttman's 1941 definition.

The procedure for computing the preliminary weights is not due to Guttman. Indeed it is unlikely that he would approve of weights calculated on the assumption that variation in the attitude is normally distributed. There are undoubtedly conditions under which such an assumption breaks down. For instance, under a two-party system favourableness towards one political party is usually correlated with unfavourableness towards the opposing party so closely that both can be scaled as a single attitude; and in the atmosphere of a general election people's attitudes may swing away from a central point of indifference towards the two extremes, and so produce a U-shaped distribution.

Nevertheless, the conditions under which the assumptions hold are very general. It is essential that each individual's attitude should be influenced by a large number of small independent causes; this will generally be true of attitudes not exposed to intensive propaganda. Moreover, the distribution of scores obtained with a given scale must tend towards normality as the number of partially independent items in it increases, whether the preliminary weights involve such an assumption or not. The break-down of the iterative process described in the previous section suggests that the attitude measured on that occasion had a skew distribution; but even this is not certain. It is arguable that more responses should have been offered for the expression of high degrees of favourableness and fewer for low, i.e., that the scale, not the sample, should be blamed for the skewness.

An efficient statistical test of the significance of differences between the weights for responses to a single item would be highly desirable. A method analogous to that of multiple regression may be found applicable to the optimum weights. These must give each response a contribution to the total variance of the scores that is always positive, however small; and the question to be decided will be whether or not its contribution is significantly greater than a residual error term. The latter will presumably be a mean square variance based on $u-n-1$ degree of freedom. If so, in a well-designed experiment u should exceed n substantially. It may prove to be a drawback of scalogram boards that they tend to encourage experimentation with samples of u about the same size as n .

Weighting Responses to Items in Attitude Scales

If the weights finally assigned to two responses do not differ significantly, it may be concluded that they are practically synonymous, and may therefore be bracketed together as a single response, with average weight. But, though this may improve the appearance of rank order, it cannot increase the precision of the scale. The responses that should be discarded altogether are those which tend to attract individuals from scattered parts of the scale of favourableness; the variance of the scores will need more consideration than the mean, and in order to secure it special tabulations or computations may be required. If its deletion is advisable, the question could be raised: which of the other responses would have been chosen in its absence by those who chose it? But in all probability those who chose it could safely be allocated to the other alternatives by comparing their scores with the averages for the responses; and the observations on the aberrant response need not disturb the estimation of the weights for the other responses from the scores of those who chose them. Thus deleting one response from an item would not entail recalculating the weights for the others.

There remains the practical problem of comparing the costs of the various procedures. In practice scalogram boards may be sufficiently quick, inexpensive, and reliable to compensate for their theoretical inadequacy, provided of course that they are available when needed. Accurate and delicate engineering is wanted to make them, and skill is needed to use them. Both imply time. It was largely for this reason that we were led to experiment with other methods and reconsider the problems involved. And it is hoped that the experience of Science 4 may be of some methodological interest, that the pencil-and-paper technique may prove helpful at least for small scale researches, and that larger organizations in a hurry for results may find some advantage in the Hollerith procedure here described.¹

¹ I am deeply indebted to my colleagues in Science 4, particularly to Dr. John Parry, who directed the researches concerned, and to Dr. Gurth Higgin and Mr. Ken Tilley, who were mainly responsible for carrying them out, constructing the scales, and collecting the data. I have also to thank the Air Ministry for permission to publish.

REFERENCES

1. Burt, Cyril (1955). 'Scale Analysis and Factor Analysis.' *Brit. J. Psychol., Stat. Sect.*, VI, 5.
2. Fisher, R. A. and Yates, F. (1943). *Statistical Tables for Biological, Agricultural and Medical Research*. London: Oliver and Boyd. (Tables XX and XXI.)
3. Garrett, H. E. (1947). *Statistics in Psychology and Education*. London: Longmans Green & Co. (Table 27).
4. Guttman, Louis (1947). 'The Cornel Technique for Scale and Intensity Analysis.' *Educ. Psychol. Measmt.*, VII, 247.
5. Horst, Paul, et al. (1941). *The Prediction of Personal Adjustment*. New York: Social Science Research Council. (Appendices by L. Guttman.)
6. Kuder, G. F. and Richardson, M. W. (1937). 'The Theory of the Estimation of Test Reliability.' *Psychometrika*, II, 151.
7. Pearson, Karl, et al. (1930). *Tables for Statisticians and Biometricians*. London: Cambridge Univ. Press. (Part II, Introduction, p. xxii.)
8. Stouffer, S. A., et al. (1950). *Social Psychology in World War II*, Vol. IV. 'Measurement and Prediction.' Princeton: Princeton Univ. Press. (Chapters by L. Guttman.)

THE STATISTICAL ANALYSIS OF THE LEARNING PROCESS

I. TASKS WITH TWO RESPONSES ONLY¹

By CYRIL BURT and ELIZABETH FOLEY

University College, London

Abstract.—The following paper deals with the application of factorial techniques to a special type of temporal variation, namely, that in which the successive measurements obtained show a progressive trend; and presents an elementary exposition, mainly from a factorial standpoint, of stochastic procedures as applied to learning and analogous processes. The formulæ and methods deduced are illustrated by experiments on simple forms of learning (a) in animals and (b) in children.

The Application of Factorial Techniques to Temporal Variations. The quantitative assessments with which the statistical psychologist is concerned can generally be specified by reference to three dimensions: each is a measurement of (i) some particular mode of behaviour ($m = 1, 2, \dots, s$ say) for (ii) some particular person ($p = 1, 2, \dots, N$) at (iii) some particular time ($t = 1, 2, \dots, n$). Accordingly, when we endeavour to express the observed data in terms of hypothetical factors, we may be concerned with three kinds of 'factor-loading'—(i) 'factor-saturations' for tests, traits or other behavioural characteristics, (ii) 'factor-measurements' for persons or groups of persons, and (iii) 'factor-weights' for the times or successive occasions at which the observations are made.² Since it is difficult to handle variations in three dimensions simultaneously, it is usual, by some form of selection or averaging, to reduce m , p , or t to 1.

In the earliest experiments on mental testing it was an almost invariable rule for every test to be applied on at least two occasions, often more. The original purpose was to eliminate, by repeating and averaging the tests, the minor errors arising from the peculiar conditions of any one particular occasion. But it was quickly discovered that while certain types of test were practically independent of time, others showed considerable changes. We thus have a division of psychological problems analogous to the distinction between 'statics' and 'kinetics' in physics.

(a) *Temporal Independence.* It was natural that the pioneers in individual psychology should deal first with problems of the former type. Here they have been fairly successful; and it has proved possible to elaborate tests and rating-methods which, with due precautions, will yield measurements that seem nearly, if not quite, free from all temporal influences. As a result, the simplified mathematical model which the factorist today employs commonly assumes that the 'abilities' estimated are stationary abilities: often, indeed, they are regarded as congenitally determined, and therefore unaffected by experience and relatively invariant in time. If after all there are variations with time, they are either ignored as of negligible importance, or treated as part of the 'unreliability' of the test. The analysis and assessment of errors of this kind formed the subject of the two papers on reliability in the last issue of this Journal.³

(b) *Temporal Dependence.* However, in some of the earlier investigations,⁴ it was observed

¹ This paper is one of the elementary 'expository notes' on particular topics which the Journal Committee has suggested should be published. It forms a sequel to the paper on reliability in the previous issue, and will be followed by others on learning and remedial education. It is intended largely for those concerned with class instruction on elementary psychology, and therefore makes little or no claim to originality in methods or results.

The senior writer is greatly indebted to his former student and colleague, Mr. A. R. Jonckheere, who has read an early draft, and offered many valuable criticisms and suggestions. Both the authors are also indebted to other research-students who have generously put their data at our disposal.

² There is no essential difference between a 'saturation', a 'measurement', or a 'weight'; and the choice of noun is chiefly the result of custom or convention.

³ C. Burt, 'Test Reliability estimated by Analysis of Variance', this *Journal*, VIII, 1955, pp. 103-18; A. F. Mahmoud, 'Test Reliability in Terms of Factor Theory', *ibid.* pp. 119-35.

⁴ Cf. C. Burt, 'Experimental Tests of General Intelligence', *Brit. J. Psychol.* III (1909), pp. 105 f.; *Distribution of Educational Abilities* (1917), p. 60. Somewhat similar difficulties arose in experimental and statistical work on psychophysical problems: at first the effects of temporal differences were regarded as part of a 'time error'; later they were studied in and for themselves.

that in certain cases a repeated application of the same or similar tests revealed a definite temporal trend. The trend was sometimes positive, and was then usually attributed either to maturation or to learning; and sometimes negative, and was then explained by boredom, fatigue, or progressive ill-health. Often such trends appeared to obey some general rule or law; and, in order to ascertain its nature, special methods of factorization were developed, popularly known as *O*- and *P*-techniques. Thus, the common criticism of factorial psychology—that it ignores the dynamic aspects of mental process—is by no means justified. Nevertheless, it must be admitted that hitherto far too little research has been carried out in this particular branch of the subject.

In dealing with such problems the advantage of a factorial technique springs from the fact that we can then postulate several temporal components. The learning curve presupposes only one. The equations of Hull, Thurstone, and W. L. Valentine assume that the learner pursues a simple path or trajectory like that of a projectile: whereas in point of fact the course followed by a delinquent who ultimately deteriorates, or a psychoneurotic patient who eventually improves, is more like that of the nocturnal motorist in one of Algernon Blackwood's stories, who tossed a coin whenever he came to a fork in the road: a branching hierarchy of alternative routes lies open before him.¹

Stochastic Processes. Here, as elsewhere in psychology, the only 'laws' that we can hope to formulate will be generalizations expressed in terms of probabilities, not of certainties. Even in dealing with physical states it has often proved impossible to formulate the laws of change in such a way that deductions can be expressed as *certain* predictions. Thus Newton's laws of motion are 'deterministic' laws: they yield an exact result. The laws of Brownian motion (the formulae describing the motions of tiny particles in a liquid) claim to be no more than 'probabilistic laws'. Partly as a result statistical theorists have during recent years devoted considerable attention to what may be loosely called the laws of probability as applied to temporal processes; and the properties of the appropriate mathematical models—many of which (as we hope to show) seem admirably suited to describe progressive changes in mental behaviour—are now well known.

The processes and laws with which we shall here be concerned may be conveniently referred to as 'stochastic'. In non-mathematical language a 'stochastic process'² may most simply be described as any process considered as developing in time and controlled by probabilistic laws. In the abstract mathematical treatment, however, the parameter values, $1, 2, \dots, t, \dots, n$ or ∞ , need have no reference to time as such: they need not even be 'real' numbers, or numbers at all. They are merely wanted to help in formulating the particular types of statistical dependence.

Mechanical Models. Karl Pearson, it is said, used to spend many of his vacations studying the methods of experienced gamblers at Monte Carlo, and trying to guess their schemes; and Professor Hobhouse, who made somewhat similar observations, reports that his experiences were 'rather like watching a not very intelligent animal trying to master a laboratory task by daily trial and error'. Certainly, to deduce the 'systems' of those *habitués* who stake solely on *rouge et noir*, or adopt one of the more popular martingales, is by no means difficult, as one of the present writers can testify; and, when the curves of their losses are plotted, it will often be found that they resemble those for the errors of successful and unsuccessful cats or rats dealing with a puzzle-box or maze.³ Such analogies are well worth exploiting. After all, the use of mechanical models to secure a practical solution for large or intractable equations is now a common expedient with mathematical physicists; and we

¹ For an ingenious stochastic method of studying such ramifying alternatives, see G. Pólya, 'Sur la promenade au hasard dans un réseau de rues', *Actualités Scientifiques et Industrielles*, 1938, no. 734.

² In current literature the term 'stochastic' is used in several different senses. In English the word has a perfectly respectable history. Jonathan Swift described himself as 'Master of the Stochastic Art' (*Precedence between Physicians and Civilians*, 1720); and we read of Sir Thomas Browne that 'though no prophet, he excelled in that faculty which comes nearest to it, namely, the Stochastic, wherein he was seldom mistaken as to future events, as well public as private' (Whitefoot, quoted in Browne, *Works*, 1712, I, xlvii). In Plato *ἡ στοχαστική* (sc. *τέχνη*) is used to designate 'the art of conjecture' (*Phileb.* 55 E). The adjective is related to *στόχος*, 'an aim or guess', and the verb *στέιχω*, 'to move towards, especially in a row or file': (cf. German, *steigen*). The Greek thus carried with it the notion of a directed temporal sequence as of an arrow or javelin in flight (*στόχασμα*). Many statistical writers seem to use the word merely as a more elegant synonym for the clumsy adjective 'probabilistic'. But of late it has acquired a somewhat narrower connotation. This may be expressed in more formal language by saying that the term denotes any family of random variables with one parameter possessing some of the characteristic properties of a temporal sequence. During the past year or so several textbooks on the subject have been available: cf. M. S. Bartlett, *Introduction to Stochastic Processes* (1955); J. L. Doob, *Stochastic Processes* (1953); and, for a more elementary treatment, W. Feller, *Introduction to Mathematical Probability* (1950), chap. XV; or Michel Loeve, *Probability Theory* (1955), p. 28 f.

³ The first to suggest the formal similarity between certain ball-and-urn problems and what Lloyd Morgan had termed 'learning by trial and error modified by chance success' was, I believe, Udney Yule, who proposed equations analogous to those developed in studies of accident-processes and of the acquisition of immunity from epidemic disease. This treatment in turn was suggested by Pearson's use of a ball-and-pigeonhole scheme for studying the incidence of cancer (*Biometrika*, IX, 1913, pp. 28 f.). Thomson developed Yule's suggestion, basing his model on a pack of cards: (if the ace of hearts represents success, then when the trainee, as it were, draws the ace of hearts, the experimenter adds a second ace of hearts to the pack: *Instinct, Intelligence, and Character*, 1924, p. 44). Thurstone revived the ball-and-urn model, and discards a black ball whenever an error is eliminated (*J. Gen. Psychol.* III, 1930, pp. 469 f.; cf. H. Gulliksen, *ibid.* XI, pp. 395 f. and *Psychometrika*, XVIII, 1953, pp. 297-307).

suggest that the same device might be profitably adopted by the educational psychologist, both for research and for class instruction.¹

In this paper we shall restrict ourselves to binary tests or trials, i.e., those permitting two alternatives only—Right or Wrong. The results can readily be represented by using balls or counters of two colours only, (say) Red and White. The probabilities will be determined by taking appropriate proportions of each colour. The counters will then be shuffled and mixed, and drawn one at a time from a bag.² The trainee will be represented by a gambler who draws a counter at random from such a bag in a game of chance where a red counter ranks as a win. We may suppose that at the start the proportions of Red and White counters (i.e., of Right and Wrong choices) are R/N and W/N respectively ($N = R + W$); and when we turn from the mechanical model to the human trainee, we can picture R as denoting the number of neural connections, excited by the stimulus, which lead to a Right response, and W the number leading to a Wrong response.

How then are we to represent the effects of practice? The simplest hypothesis would be to suppose that every trial produces some effect and that the effect is always the same—say, adding D more neural connections leading to the Right response. Then, after t trials, y , the probability of a Right response, will be $(R + tD)/(N + tD)$. The curve so described will evidently be a rectangular hyperbola—a curve which has in point of fact been suggested by several investigators on other grounds.

In most of our experiments, however, the effect will vary according as the trainee succeeds or fails. In the mechanical model this may be represented by supposing that, when the player draws a counter, then, instead of simply replacing it, $1 + D_r$ red counters are returned to the bag after a red counter has been drawn, and $1 + D_w$ white counters after a white has been drawn: in the former case the probability will therefore change from R/N to $(R + D_r)/(N + D_r)$ and in the latter to $R/(N + D_w)$. With the classical procedure, where counter or ball is always supposed to be replaced, $D_r = D_w = 0$. With the less usual procedure,³ where it is not replaced but no further change is made, $D_r = D_w = -1$. With the more general procedure with which we are here concerned, the two parameters, D_r and D_w , will have to be estimated from the data.

There are, however, many conceivable cases where this model will still not suffice: e.g., those in which failure is practically certain at the first trial. In such cases we must imagine that counters of both colours are added whether the trainee fails or succeeds. This yields four parameters, R_r , R_w , W_r , W_w (say), one or more of which may be negative. Finally we must allow for the possibility that the effects progressively change: so that the numbers added will be p_1R_r , p_2R_r , ..., etc., with similar values for the other eventualities.⁴ In that case we naturally look for some kind of interdependence between the successive values of p .

Classification of Stochastic Processes. With one or other of these schemes we have found it possible to reproduce a plausible imitation of practically any set of results that we have obtained experimentally. But it is scarcely practicable to use them for purposes of systematic analysis unless some kind of classification is introduced. The commonest method of classification begins by

¹ The ordinary statistical textbook leads students to suppose that the only reputable way to cope with a numerical problem is to derive a functional equation by theoretical analysis, and solve that. In what is rather picturesquely called 'the Monte Carlo method' the physicist reverses the procedure: he converts his functional equation into a problem in probabilities; and by setting up an urn-and-ball model or turning to a table of random numbers he obtains an approximate solution in a quicker and more practical fashion. Readers of Todhunter's *History of Probability* (1865) will recall how many theorems contributed by Galileo, Pascal, and Fermat were first suggested by inquiries from eminent gamblers; 'Student' originally obtained the sampling distribution of the correlation coefficient in the same fashion; and half a century ago Lord Rayleigh and Einstein drew attention to the practical value of such devices in solving physical problems.

A Symposium on 'the Monte Carlo Method' was held at Los Angeles in 1949; cf., Nat. Bur. Standards, U.S.A. 'Monte Carlo Method', *App. Math. Ser.*, xii (1951). A briefer summary of the subject, with references, will be found in the Appendix to W. E. Milne's *Numerical Solution of Differential Equations* (1953). For the use of random numbers in solving probability-problems in the theory of learning, see R. R. Bush and F. Mosteller, *Ann. Math. Statist.*, XXIV (1953), pp. 559-85.

² In most continental resorts, where there is a popular casino, moderately reliable apparatus can usually be purchased for experiments of this kind. We ourselves have used beads, such as are supplied for kindergarten work; but it is essential that they should have the same weight as well as the same size and surface-texture. For mixing we have adopted a device similar to that used by Galton and Pearson—a sloping board studded with stout pins, ending in a funnel: (cf. F. Galton, *Natural Inheritance*, 1889, p. 63, fig. 9). For group experiments we give a bag to each of a number of pupils. The results so obtained are particularly helpful in the study of sampling distributions and methods of estimation. Our most striking discovery was the wildly erroneous conclusions furnished by the chi-squared test.

³ This is the most general case explicitly considered in Pearson's system of frequency curves: (cf., W. P. Elderton, *Frequency-Curves and Correlation*, 3rd edn., 1938, pp. 39 f.).

⁴ This general formulation is essentially that of G. Pólya in his paper 'Sur quelques points de la théorie des probabilités', *Ann. Inst. H. Poincaré*, I (1931), pp. 117 f. Sect. II, 'Sur quelques courbes de fréquence résultant de la superposition d'effets fortuits interdépendants'. He applies his formulæ to the transmission of hereditary characteristics and the spread of contagious disease—illustrations which have provided later writers with a metaphorical terminology. Similar probability models were proposed by Léon Brillouin to describe hypotheses in quantum-theory (*Ann. de Physique*, VII, 1927, pp. 315-31).

More recently Bush and Mosteller have constructed artificial mathematical models for learning processes by using tables of random numbers, and christened them 'stat-rats' (*Ann. Math. Statist.*, XXIV, 1953, pp. 559 f.). Miss Foley has made analogous experiments with bags and beads, and one of her schemes yields an excellent fit to the curve for cats in Table I; accordingly she has dubbed her mechanical model a 'stat-cat'.

dividing such schemes into two main types—non-Markov and Markov,¹ or, as Pólya terms them, *globales* and *liées en chaîne*. In the former case the result of the $(t+1)$ th trial is dependent on the results of each of the t preceding trials; in the latter case it is dependent solely on that of a single trial, namely, its immediate predecessor. On factorizing the data, the former hypothesis would be represented either by a general factor or (if we prefer) by a broad group factor, and the latter by a specific. Any further classification can, in our view, best be achieved by introducing supplementary factors. Thus, in psychological work it is frequently necessary to introduce narrower group-factors. Of many learning processes, for example, it is tempting to echo what Madame du Deffand said of the famous walk of the decapitated saint, *ce sont les premiers pas qui coûtent*; and of others it would be truer to say that the more recent trials are also the most influential. Similar factors also distinguish what seem to be the effects of delayed memory from those of immediate memory.

To cover all conceivable possibilities a fairly flexible technique is plainly needed. We suggest that the time-range of the learning process should be regarded as covering a finite number of trials, starting with the prior state of unstable equilibrium, which the first trial disturbs, and ending with the final state of stable equilibrium, when learning is complete. Suppose, for example, we are dealing with a single task, and that the observed data form a matrix of measurements for n trials or 'occasions': we can apply *O*-technique, correlate or covariate the trials, and then factorize the matrix of product sums. Each temporal factor will cover the whole time-range; and for n trials there will be n factors. But, unless we find it necessary to assume foresight or conscious purpose, the factor-saturations for the r th trial will be zero for the $(t+1)$ th to the n th factors, and the factor-matrix will thus be triangular. The remaining saturations will give varying weights to the different antecedent trials. A rotation of factors may then prove desirable, leading usually to both broad and narrow group factors, and, if the process is predominantly of the Markov type, to large specific factors.

So-called *O*-technique, however, is both unfamiliar and decidedly onerous. We shall accordingly begin with a much simpler procedure, in which, as is more usual, tests or trait-measurements will be correlated; but the correlation will be carried not over the various persons, but over the various trials or 'occasions'. And we shall start by supposing that a separate analysis is to be made for each separate trial.

Basic Formulae. In experiments of the kind we propose to report, a sample of N individuals are subjected, at each of n trials, to a set of u stimuli or test-items, each stimulus being capable of provoking one of u alternative responses. Hence at each trial there may be in all $u \times v$ different responses or u^v different patterns of response: in general, let us say, s possible modes of behaviour. Accordingly, the observed performance of the child will be entered in an observation matrix, M , of order $s \times N$.

Before the first trial each child's mind is defined (we may suppose) by r 'abilities'. These we can regard as hypothetical quantities specifying the state of his mind at time 1. They will thus form an $r \times N$ matrix which (following the usual notation) we will call P . The effect of the tests on the s possible responses will be indicated by an $s \times r$ operator, F , analogous to a matrix of 'factor-saturations' or 'factor-loadings'. If (as with the simplest forms of solution) there are as many factors as tests, then $r = s$, and F will be a square matrix.

To express the outcome of the first application in terms of this hypothesis we write,

$$F_1 P_1 = M_1 \quad (i)$$

which in form is identical with the 'fundamental equation' of factor analysis. For many purposes it is useful to normalize the columns of F , and put

$$M_1 = F_1 P_1 = L_1 V_1^{-1} P_1, \quad (ii)$$

where V_1 is a diagonal matrix.

Were our aim merely to assess the 'reliability' of the tests, the same set (or a parallel set) would be applied on a second occasion; and if the results are unaffected by previous experience or by the events of the intervening period, then the only difference discernible between the first set of matrices and the second should consist of random variations obeying what are loosely called the laws of chance. But if several months of growth or education elapse between the two applications, there will generally be significant differences between M_1 and M_2 , V_1^{-1} and V_2^{-1} , and P_1 and P_2 ; but (so it is usually found) little or no significant difference between L_1 and L_2 .²

In this and the following paper we shall be concerned chiefly with educable abilities. In such a case we can no longer assume that the successive measurements are independent. As a result of the first trial (or the events that followed) we shall therefore suppose that the 'abilities' change from

¹ The problems raised by interdependent trials were first systematically studied by A. A. Markov, who drew special attention to what he called 'verketteten Ereignisse', i.e., series in which each event is directly linked only to the last: (cf., 'Untersuchung eines wichtigen Falles abhängiger Proben' (in Russian), *Abh. der K. Russ. Ak. der Wissenschaft* (1907) and other papers; and, for a general summary, *ibid.*, *Wahrscheinlichkeitsrechnung* (trans. from 2nd Russian ed. by H. Liebmann) (1912), *Anhang II*, pp. 272-98).

² Cf., C. Burt, 'The Differentiation of Intellectual Ability', *Brit. J. Educ. Psychol.* XXIV (1954), p. 76, esp. refs.

P_1 to P_2 ; and we shall express this influence by a linear operator, W_1 . Assuming as before that linear relations will suffice to describe the effects, we may accordingly write,

$$P_2 = W_1 M_1 = W_1 L_1 V_1^{\frac{1}{2}} P_1. \quad (\text{iii})$$

The observable effect of the second application of tests will then be expressed by

$$F_2 P_2 = M_2. \quad (\text{iv})$$

If after all M_2 is not significantly different from M_1 , we shall have, to a close approximation, $M_2 = M_1 = L_1 V_1^{\frac{1}{2}} P_1$, and $P_1 = V_1^{-\frac{1}{2}} L_1^{-1} M_1$; so that

$$W_1 = V_1^{-\frac{1}{2}} L_1^{-1}, \quad (\text{v})$$

and $P_2 = P_1$. It will be seen that eqn. (v) is analogous to the equation used in factor analysis by weighted summation to derive the matrix of 'weights' from the matrix of 'saturation'. But in the present case we do *not* assume that L is an orthogonal matrix—merely that it is non-singular. If however M_1 and M_2 exhibit significant differences, then (as we shall discover later on) the differences can often be accounted for almost entirely in terms of the diagonal matrix $V^{\frac{1}{2}}$.

Substituting in eqn. (iv) the value for P_1 given by eqn. (iii), we have

$$M_2 = F_2 W_1 M_1 = T_1 M_1 \text{ (say);} \quad (\text{vi})$$

and, if $F_2 = L_1 V_2^{\frac{1}{2}}$ and $W_1 = V_1^{-\frac{1}{2}} L_1^{-1}$, then

$$T_1 = L_1 V_2^{\frac{1}{2}} V_1^{-\frac{1}{2}} L^{-1} = L_1 D_1 L^{-1} \text{ (say),} \quad (\text{vii})$$

where D_1 is a diagonal matrix. We can thus (as in so many applications of factorial theory) often dispense with an explicit calculation of P_1 . Later trials will generate similar equations.

In the present paper, in order to start with the simplest possible model, we shall assume that the effect of each successive trial is approximately the same, i.e., that $L_1 = L_2 = \dots = L_{n-1} = L$ (say) and $V_1 = V_2 = \dots = V_{n-1} = V$ (say). Hence the successive values of F , W , and T will also be the same. We shall show that, for many experiments in which either children or animals have been subjected to a course of training, this assumption of constancy will yield a plausible fit to the actual data. Accordingly, assuming T to be constant, we may write

$$M_n = T M_{n-1} = T T M_{n-2} = \dots = T^n M_1 \quad (\text{viii})$$

$$= L D L^{-1} \cdot L D L^{-1} \dots L D L^{-1} M_1 = L D^n L^{-1} M_1. \quad (\text{ix})$$

Eqns. (viii) and (ix) really supply us with $n-1$ distinct sets of equations for evaluating T ; and the several sets will not be completely consistent. If, however, we adopt the principle of least squares, we can find the values that yield the closest fit by following the methods adopted for solving multiple regression equations. Eqn. (ix) will evidently be satisfied if we take D and L to be the latent roots and latent vectors (or 'modal columns') of the square matrix T . Convenient computing methods are available for factorizing T in this way.¹

We can of course evaluate each of the matrices, $T_1, T_2, \dots, T_{n-1}$, separately, and so determine values for $L_1, L_2, \dots, L_{n-1}$, and for $V_1, V_2, \dots, V_{n-1}$. Then, if they are not approximately constant, it may still be possible to discern a definite trend. Occasionally, for example, a better fit may be secured by writing $T_2 = U T_1$, $T_3 = U T_2 = U^2 T_1$, and so on; in short, by treating T as we have already treated M . But these further elaborations will not concern us here.

It will be seen that the effects of 'learning' are described essentially by the matrix T . Such a description involves no particular theory of learning; we can regard the changes either as due to the cumulative strengthening of associations (in accordance with the theory of temporal contiguity), or as due to reinforcement (in accordance with the so-called law of effect), or as the result of intelligent insight. To this extent the factorial model is more general than any of the existing theories. However, by introducing appropriate modifications into the conditions of the experiment it should be possible to design investigations which may decide between them. In point of fact, these rival modes of explanation are by no means mutually incompatible; and, so far as our own experiments go, each of the three principles appears to be operative among pupils of school age, their relative importance varying chiefly with the mental age of the children and the nature of the task. When sufficiently large samples are available and the full factorial procedure is applied, we have often found as many as three significant factors; and from indirect evidence it has at times appeared plausible to identify them roughly with these three types of causative process. Judged by their respective contributions to the total variance, it would seem that their relative predominance alters with increasing age and increasing intelligence, that of the first diminishing and that of the last increasing.

Nevertheless, the statistical mode of approach here suggested need not of itself imply that prior events in the temporal sequence are to be regarded essentially as *causes* of subsequent events. In the

¹ Cf., R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices* (1938), pp. 142 f.

first instance the prior events will be treated merely as furnishing antecedent *information* by the aid of which the subsequent events can be *predicted* with a greater or lesser degree of likelihood. Thus it is better to expound the underlying mathematical arguments in terms of information theory than in terms of some specific causal theory. Primarily, what is supplied by the factor theory is no more than a simplified mathematical model for analysing certain data. It is a distinct and separate step to identify the abstract elements of the model with the concrete elements of the situation which it is intended to represent.¹

Pass-or-Fail Tests. The assumptions and inferences underlying the foregoing algebraic arguments will become clearer if we apply them to the simplest conceivable type of problem. With many psychological tests only two modes of response are possible, and these are mutually exclusive and jointly exhaustive, e.g., responses that can be classed as either definitely right or definitely wrong, in short, as either successes or failures. The items that comprise the Binet scale and many of the tests devised to measure educational attainments are of this nature. Usually with any one item, if we take a batch of children sufficiently young, we shall find (when guessing is excluded) that every child at first fails, and that, after a sufficient interval of growth or learning, every child passes. If we use 0 to denote failure and 1 to denote success, then the matrix M , the initial table of results, will consist of an 'answer pattern' containing nothing but units and zeros. Such a matrix can be subjected to a factorial analysis by the usual techniques just like any other measurement matrix.²

In dealing with the theoretical problems of learning, however, we are more frequently interested in the general course of progress shown by the average child than in the progress of individuals. Hence for each successive trial we can often sum together the results obtained by the sample as a whole; and thus greatly simplify the mathematical analysis. In such cases the entries in the initial matrix are usually expressed in the form of percentages. For example, in an early investigation with the Binet tests we discovered that at the age of 4 no child can repeat the months of the year; at 6, 7 per cent. can do so; at 8, 62 per cent.; at 10, 93 per cent.; at 12, 100 per cent.³ We can, if we like, think of these percentages as stating the *probability* that, at this or that age, any individual child picked at random will answer this problem correctly.

However, with test-items composed of meaningful material the child's progress is liable to be affected by a wide variety of influences, many of them not amenable to experimental control. Hence it will be better to start with tasks which from their very nature are not likely to be influenced by extraneous conditions. To secure illustrative examples that would demonstrate the practical possibilities of the methods proposed we have searched the records available from our own work and that of our colleagues, and have selected the two which show the smallest amount of erratic disturbance.

Example 1. We first take a simple experiment on conditioning. For purposes of popular exposition it has the additional merit of being at once less artificial and more easily pictured than the more carefully devised experiments of the laboratory; and its very crudity suggests instructive points for class discussion and criticism. The data we owe to Mrs. E. N. Phillips, a former teacher who has become a successful breeder of Siamese cats. We may describe the procedure in her own words:

"I train them early to respond to two distinct calls: a bicycle bell, which means "Move away from there at once", and a police whistle, which means "Come here (for your meals or the like)". In the first case the 'conditioned stimulus' (as the psychologist would call it) is the bell; and the 'unconditioned stimulus' three or four drops of water squirted on to the back after an interval of 3 seconds, if the animal does not respond to the bell. Some of the creatures were decidedly slow in learning to obey the signal; but once they had responded to it they usually continued to do so at later trials without the need for any further stimulation. All were practically perfect after about 20 trials; but nearly every one was liable to an occasional lapse, no matter how long the training was continued. The differences between individuals, however, were not so marked as in the training with the whistle; and the correlation between the two, though positive, was low. For the bell experiment I give figures for 28 of the animals: with three the training was a little irregular owing to inter-current illness."

For purposes of illustration we want data that will run as smoothly as possible. Hence we have here omitted the three most erratic animals. For the remaining 25 the figures for the first 21 trials are as follows:

¹ This warning is especially necessary in discussing the theory of 'factors', since the student is apt to suppose that the very word refers to causal agencies. Indeed, Spearman and his followers insisted on defining the term in that sense. In the mathematical analysis they are of course merely 'components'.

² Cf., C. Burt, 'The Factorial Analysis of Qualitative Data', this *Journal*, III (1950), pp. 171 f.

³ C. Burt, *Mental and Scholastic Tests* (1921), Table III, pp. 132-33. There are undoubtedly considerable risks in dealing with the pooled results from the group as a whole instead of with the results for each individual taken separately; a probability based on an average of frequencies is not always the same as one based on an average of the corresponding probabilities. But this is a question to which we shall return later.

TABLE I. EXPERIMENTS ON TRAINING (1)

Correct Responses at Each Trial

Trial (t)	1	2	3	4	5	6	7	8	9	10	11
Frequencies Observed Estimated	0 0.0	3 3.7	6 6.8	7 9.4	11 11.6	16 13.5	13 15.1	15 16.5	14 17.6	18 18.6	16 19.4
Proportions Observed Estimated	.00 .00	.12 .15	.24 .27	.28 .38	.44 .46	.64 .54	.52 .60	.60 .66	.56 .70	.82 .74	.64 .78

Trial (t)	12	13	14	15	16	17	18	19	20	21
Frequencies Observed Estimated	20 20.1	19 20.6	22 21.1	21 21.5	23 21.9	23 21.9	25 22.2	24 22.7	25 22.9	24 23.0
Proportions Observed Estimated	.80 .80	.76 .82	.88 .84	.84 .86	.92 .88	.92 .89	1.00 .90	.96 .91	1.00 .92	.96 .92

British psychologists commonly describe the effects of such training by saying that, as a result of experience, certain tendencies or 'dispositions' to respond in this or that way are gradually formed in the mind or central nervous system.¹ Thus the fractional entries in the above table might be regarded as describing the strength of such a 'disposition' as it is progressively built up. The value .80, for example (shown for the 12th trial) may be interpreted as stating that the probability that a typical animal will respond correctly on this occasion is .80; or, if we are dealing with a group, that 4 out of 5 animals in a reasonably large and random sample will do so. It thus describes the hypothetical state of a typical creature's mind just before this phase of the experiment.

The Postulate of Non-mnemic Causation. Now a useful simplification, adopted in most physical sciences as a working hypothesis, consists in the postulate that all causation is temporally contiguous. This is equivalent to a denial of what has been termed 'mnemic' or 'transilient' causation. It is expressed by the old scholastic adage: *Natura non agit per saltum*. The most celebrated formulation of the principle is the oft-quoted pronouncement made by Laplace at the outset of his *Essai philosophique sur les probabilités* (1814): 'We are to consider the present state of the universe as the effect solely of the immediately preceding state and as the cause of the state that is to follow'. Thus in the Newtonian scheme of dynamics certain differential equations can be found in virtue of which, given the configuration and the velocities of a material system at any one instant of time, it is theoretically possible to predict the configuration of the system at the next, and therefore at any later, instant. A knowledge of its earlier history is thus irrelevant and superfluous. Events so determined are said to be 'chained': each state of the system is assumed to be linked directly to the next state, and only indirectly to other later (or earlier) states. It is this hypothesis that lends special importance to the so-called Markov process.²

In psychology a similar type of causal dependence is generally assumed. Here too, it is commonly supposed, there is no such thing as action at a distance of time. Bertrand Russell, it is true, for a while maintained that the laws of psychology are distinguished by mnemic causation—'a special kind of causation in which there is an interval of time between the cause and the effect without any intervening chain of causes'. This would certainly enable us to dispense with the hypothesis of persisting and modifiable psychical or physiological 'dispositions'; otherwise, in the form in which Russell states it, the supposition has little to commend it.³

¹ W. McDougall, *Psychology: the Study of Behaviour* (1912), pp. 66 f.; J. Ward, *Principles of Psychology* (1918), pp. 66 f.; G. F. Stout and C. A. Mace, *Manual of Psychology* (1938), pp. 313 f.

² One of Einstein's earliest achievements was to apply this procedure to deducing statistical equations for Brownian movements (*Ann. d. Physik*, XVII, 1905, pp. 549 f.). Each particle may be supposed to perform a 'random walk', and we thus eventually reach equations determining the approximate size of the molecules and their number per unit volume.

³ B. Russell, *Analysis of Matter* (1921), pp. 78 f. and 307 f. The theory is subjected to a lucid examination in C. D. Broad's *Mind and its Place in Nature*, 3rd imp. (1937), pp. 440 f. Russell has since abandoned the hypothesis as 'too violent': cf. *The Philosophy of Bertrand Russell* (1946), 'Reply to Criticisms', p. 700.

But in any case, until more empirical evidence is available from neurological work, it will be wiser not to concern ourselves with the *states* (which in psychology are never directly observable), but merely with the *behaviour* of the system, i.e., with the linking of events; and, as in modern physics, the concept of causation can be dropped as needlessly metaphysical. We may then be forced to assume that, even if prior *states* have no causal importance, nevertheless prior *events* may supply additional information of considerable value. To begin with, however, we shall use them merely for the preliminary business of determining the general rule or law.

Now, in experiments such as the one reported above, each event consists simply in selecting one of two possible responses. The selections occur in a specified time-order; and the successive selections can no longer be regarded as wholly independent, as in the more familiar problems of statistical psychology: instead we assume that each is directly dependent on the last alone. Hence, in order to calculate the probabilities of a right or wrong response on the $(t+1)$ th occasion the only information needed will be that given by the results of the t th trial, together with a general rule or law stating how the two are related.

Notation. To express these probabilities we shall adopt the notation familiarized by Jeffreys, Aitken, and others. Let us designate the two possible responses—the ‘right’ response (animal obeys the bell) and the ‘wrong’ (animal ignores the bell)—by the letters R and W respectively. Then the probability that it will make a right response on the t th trial can be denoted by $p_t(R)$. Evidently $p_t(W) = 1 - p_t(R)$, since one or the other type of response must occur. When we take into account the result of the preceding trial, we can specify the component probabilities more precisely: if at the $(t-1)$ th trial a right response was made, then at the next trial, the t th, the probability that the response will also be right may be symbolized by $p_t(R|R)$; if at the $(t-1)$ th trial a wrong response was made, then the probability that the next response will be right may be symbolized by $p_t(R|W)$. These further constants, which express the way each successive event depends on the preceding, may be termed ‘transitional probabilities’. If, as before, we use m_t to denote the mode of response, whether right or wrong, then we may formally define a Markov process by the condition

$$p(m_{t+1}|m_t, m_{t-1}, \dots, m_2, m_1) = p(m_{t+1}|m_t)$$

where, in general, $m_1, \dots, m_{t-1}$ will be vectors. Since there will be similar equations for $p(m_t|\dots)$, $p(m_{t-1}|\dots)$, etc., it follows that

$$p(m_{t+1}|m_t, m_{t-1}, \dots, m_2, m_1) = p(m_1) \prod_{j=1}^{t+1} p(m_{j+1}|m_j),$$

where $p(m_{j+1}|m_j)$ is a transitional probability, i.e., a conditional probability referring to time.

We may now introduce our last simplification. Until the data force us to fall back on some more complex hypothesis, we shall assume that the transitional probabilities remain the same throughout the whole course of training. That being so, we can now drop the subscript t . To construct a Markov chain of this simple type (two alternatives only, with constant transitional probabilities), it is sufficient to know three quantities: the initial probability, $p_1(R)$, i.e., the probability of a right response at the very first trial, and two transitional probabilities, say $p(R|W)$ and $p(W|R)$.¹

Mechanical Model. We can put our mathematical model into practical form as follows. We shall use three bags; and in order to distinguish them, we shall suppose them to be made of Black, White, and Red material respectively. They will contain Red and White counters in the proportions indicated by the three hypothetical probabilities, viz., $p_1(R)$ in the Black bag, $p(R|W)$ in the White bag, and $p(W|R)$ in the Red. Each of the 25 trainees, on entering the laboratory, goes first to the black bag, draws a counter, shows it to the experimenter, and returns it at once to the bag. According to the colour he drew, the experimenter now directs him to either the white bag or the red. Here he draws another counter, which the experimenter replaces by another of the same colour, ready for the next trainee. When all have drawn their second counter, they go back to the black bag, which has meanwhile been emptied, and drop their counters into it. The experimenter counts the contents, and the proportions found, $p_2(R)$, and $p_2(W)$, are recorded as showing the probabilities of a right or a wrong response at the second trial. The trainees now start all over again, drawing from the first bag, and being directed once again, according to the colour they have drawn, to one or other of the second pair.

To obtain a series of results like those in Table I, the trainees begin by drawing from a bag containing nothing but white counters, since $p_1(R)$ here = 0; and the trials are repeated until its new proportions cease to vary significantly, i.e., until practically all the counters are Red.

Estimation of Transitional Probabilities. Reverting to the actual data obtained from the experiment with cats, our first problem is to decide how the transitional probabilities may best be estimated. Their function is to enable us to predict the nature of the $(t+1)$ th response from that of the t th. The

¹ The use of Markov processes in the theory of learning has been discussed by G. A. Miller (*Psychometrika*, XVII, 1952, pp. 149-67) and several other writers. For a more general treatment see M. Fréchet, ‘Recherches théoriques modernes sur la théorie des probabilités’, *Traité du Calcul des Probabilités et ses applications* (ed. E. Borel), I, no. iii, 1937-8.

most natural procedure therefore will be to determine the correlation between the two. Here for theoretical reasons it will be better merely to compute the product-sums, without reducing the correlated variables either to deviation form or to standard measure, i.e., to use what are sometimes called 'unadjusted' and 'unstandardized correlations'. Were we concerned with the training of one person only, the technique would plainly be a special form of 'P-technique', namely, correlating the data for *persons* obtained on successive *occasions*; and its peculiar nature would simply turn on the fact that the two 'persons' to be correlated were two sets of data for the same person. If, therefore, a technical label is required for the method, in order to classify it in terms of the nomenclature commonly used in this country, it may be said that we are in effect using 'P₀-technique for a single person'.¹

In the present experiment we are concerned with the effects of training not on a single individual, but on a group of 25 individuals, who are virtually treated as forming one composite 'person'. Furthermore, the two sets of figures to be correlated are also the same: the only difference is that in the second series the figures are moved up by a single step. We have thus to calculate a particular form of the 'autocorrelation function'—a function which has certain distinctive properties of its own.²

Expressed algebraically the method is a simplified instance of the procedure described in more general terms at the outset. M_t will here be the $2 \times (n-1)$ matrix of observed frequencies, for the 1st up to the $(n-1)$ th trial; and what we have to estimate is an analogous matrix for the frequencies of the 2nd to the n th trials. Let us call it $\bar{M}_{t+1}$, where $\bar{M}_{t+1} = TM_t$. We demand that $\bar{M}_{t+1}$ shall provide as close a fit as possible to the actual frequencies for those trials, i.e., to the matrix M_{t+1} . To do this we must take

$$T = M_{t+1}M_t'(M_tM_t')^{-1} \quad (x)$$

where T will be the 2×2 matrix of transitional probabilities

$$\begin{bmatrix} p(R|R) & p(R|W) \\ p(W|R) & p(W|W) \end{bmatrix}.$$

The working instructions are briefly as follows. Write down in two parallel columns the total number of correct and incorrect responses for trials 1 to $n-1$ and at the side the corresponding numbers for trials 2 to n . For reference these are here labelled a, b, c, d . Now calculate the square sums and product sums indicated by $\Sigma a^2, \Sigma b^2, \Sigma ab, \Sigma ca, \Sigma cb, \Sigma da, \Sigma db$.

TABLE II. CALCULATION OF SQUARE-SUMS AND PRODUCT-SUMS

No. of Trial	Variables				Squares		Products				
	t th trial	$(t+1)$ th trial									
t	a	$\begin{cases} b= \\ 25-a \end{cases}$	c	$\begin{cases} d= \\ 25-c \end{cases}$	a^2	b^2	ab	ca	cb	da	db
1	0	25	3	22	0	625	0	0	75	0	550
2	3	22	6	19	9	484	66	18	132	57	418
3	6	19	7	18	36	361	114	42	133	108	342
...	...	...	...	...	...	...	...	...	...	...	...
20	25	0	24	1	625	0	0	600	0	25	0
21	(24)	(1)	—	—	—	—	—	—	—	—	—
Total	—	—	—	—	6211	2661	1814	6433	2191	1592	2283

There are convenient checks on the arithmetic. With $N = 25$ and $n-1 = 20$, $\Sigma a^2 + \Sigma b^2 + 2\Sigma ab = \Sigma ca + \Sigma cb + \Sigma da + \Sigma db = 25^2 \times 20 = 12500$.

we now obtain $M_{t+1}M_t' = \begin{bmatrix} \Sigma ca & \Sigma cb \\ \Sigma da & \Sigma db \end{bmatrix} = \begin{bmatrix} 6433 & 2192 \\ 1592 & 2283 \end{bmatrix},$

and $M_tM_t' = \begin{bmatrix} \Sigma a^2 & \Sigma ab \\ \Sigma ab & \Sigma b^2 \end{bmatrix} = \begin{bmatrix} 6211 & 1814 \\ 1814 & 2661 \end{bmatrix}.$

¹ For the general nature of this procedure, see C. Burt and H. Watson, 'Factor Analysis of Assessments for a Single Person', this *Journal*, IV (1951), pp. 179-91 and refs. For the characteristics of the 'autocorrelation function', see M. G. Kendall, *Advanced Theory of Statistics*, II, pp. 421 f.

² Cf., C. Burt, *Marks of Examiners* (1936), p. 301, eqn. (xxxvii) or G. Thomson, *The Factorial Analysis of Human Ability* (1939), p. 303, eqn. 30.

The Statistical Analysis of the Learning Process

The inverse of $M_i M_i'$ is $\frac{1}{6211 \times 2661 - 1814^2} \begin{bmatrix} 2661 & -1814 \\ -1814 & 6211 \end{bmatrix}$.

Hence (dropping the last three digits throughout)

$$T = \frac{1}{13237} \begin{bmatrix} 13142 & 1945 \\ 95 & 11292 \end{bmatrix} = \begin{bmatrix} .993 & .147 \\ .007 & .853 \end{bmatrix}.$$

As a check on the working it will be observed that each column in the matrix thus computed sums to 1.000.

The initial probabilities must here be decided by *a priori* considerations. Before training has started we should in this case not expect any animal to respond correctly to the sound of the bell. The initial probabilities will therefore be $p_1(R) = 0$ and $p_1(W) = 1$, or in terms of frequency, 0 and 25 respectively.

Calculation of Expected Frequencies. We can now build up the expected numbers by repeated multiplication. We proceed as shown in Table III.

TABLE III. CALCULATION OF EXPECTED FREQUENCIES

Trial	1	2	3	4	
Number Right (<i>a</i>)	0	3.675	6.784	9.415	...
Number Wrong (<i>b</i>)	25	21.325	18.216	15.585	...
Transitional Probabilities					
.993	.000	3.649	6.737	...	...
.147	3.675	3.135	2.678	...	...
Number Right (<i>a</i>)	3.675	6.784	9.415	...	...
.007	.000	.026	.047	...	...
.853	21.325	18.190	15.538	...	...
Number Wrong (<i>b</i>)	21.325	18.216	15.585	...	...
Total (<i>a</i> + <i>b</i>)	25.000	25.000	25.000	...	...

Of the 25 who responded wrongly at the first trial 14.7 per cent., i.e., 3.675, will respond correctly at the second trial, and the rest, 21.325, will continue to respond incorrectly. Of the former 99.3 per cent. will continue to respond correctly at the third trial, namely, 3.649; 14.7 per cent., however, of those who still responded wrongly at the second trial will now respond correctly at the third, namely, 3.135. The total number of those now responding correctly will therefore be 3.649 + 3.135 = 6.784. And so on: (cf., top row of Table III and 2nd row of Table I).¹

The Matrix of Transitional Probabilities. The structure of the matrix T is of interest in itself. Four conditions call for special comment. (i) If $p(R|R) = p(W|W) = \frac{1}{2}$, then $p(W|R) = p(R|W) = \frac{1}{2}$, and the outcome of each trial will be independent of the outcome of the previous trial. (ii) If $1 > p(R|R) = p(W|W) > \frac{1}{2}$, then there will be a tendency to oscillate from a correct response to an incorrect—a tendency which is evident in many neurotic conditions. (iii) If as here $p(R|R) > p(W|W)$, then $p(R|W) > p(W|R)$, i.e., the very fact that a response is successful will increase the probability of successful responses in subsequent trials—a tendency which (as is well known) governs the formation of good and bad habits alike. (iv) Finally, unless $p(W|R)$ is zero (or is eventually reduced to zero by inhibition or some similar cause), $p_\infty(R)$, the limiting probability of a correct response, can never reach unity.

The Curve of Learning. The calculated figures, showing how the number of successes increases from each trial to the next, can be plotted in the form of a smoothed curve, which, in the present case,

¹ Incidentally the student will find it instructive to check the value of T from these theoretical values by means of the equation given above. Taking three values of the variables only, as given in Table III, he will find $\sum a^2 = 59.529$, $\sum b^2 = 1411.579$, $\sum ab = 201.946$, $\sum ca = 88.802$, $\sum cb = 408.048$, $\sum da = 172.673$, and $\sum db = 1205.477$. This gives for T

$$\begin{bmatrix} 42947.377 & 6357.388 \\ 300.158 & 36890.147 \end{bmatrix} \div 43247.535 = \begin{bmatrix} .993 & .147 \\ .007 & .853 \end{bmatrix}.$$

Note as a further check, that each column in the first matrix adds up to the determinant used as a divisor, viz., 43247.535.

yields a tolerably good fit to the original data. And it is natural to inquire how such a curve compares with the 'curves of learning' described by the various formulae, both rational and empirical, that have been proposed by previous investigators.¹

We do not wish to imply that a single equation of this kind can serve as an adequate substitute for the more detailed expressions derived from stochastic considerations. To say that, at such-and-such a stage of training, there is such-and-such a probability that a boy's reaction will be so-and-so after failure but so-and-so after success, is far more informative than saying that the general trend of his progress will be a hyperbola, or an exponential curve, with such-and-such components; and, when we have to deal with multiple responses instead of only two, the superiority of the stochastic approach is plain beyond all question. Nevertheless, in order to present a succinct summary of the final results and secure a convenient short-cut for practical computations, it will be helpful to devise a single formula (if we can) which may enable us to deduce the expected values directly from the data instead of building them up step by step after the manner of Table III.

Such an equation can be derived by factorizing T into its latent roots and latent vectors as indicated by eqn. (vii). With a matrix of a higher order, such as would be needed for a larger number of alternative responses, we should proceed by weighted summation. With a 2×2 matrix a direct algebraic solution can readily be reached. Following the usual procedure we find for the diagonal matrix of latent roots

$$D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & p(R|R) - p(R|W) \end{bmatrix},$$

and for the latent vectors $L = \begin{bmatrix} +1 & +1 \\ -p(W|R)p(R|W) & -1 \end{bmatrix}$. Moreover $M_1 = \begin{bmatrix} p_1(R) \\ p_1(W) \end{bmatrix}$. Substituting

these values in eqn. (ix) we at once obtain the formula for constructing the theoretical curve, namely,

$$p_t(R) = T^t M_1 = \frac{p(R|W)}{p(R|W) + p(W|R)} \left\{ 1 - \left[p_1(W) - \frac{p(R|W)}{p(W|R)} p_1(R) \right] \left[p(R|R) - p(R|W) \right]^t \right\} \quad (\text{xi})$$

where t as an index takes the values 0 to $n-1$. (In the tables the trials have been numbered in the more natural way, calling the first trial 1; for purposes of the following formulae, however, the index of the trial which involves no previous practice must be given the value 0, and the subsequent indices numbered accordingly.)

As t increases indefinitely, the expression $[p(R|R) - p(R|W)]^t$ becomes increasingly small; so that the value of $p_t(R)$ converges to $\frac{p(R|W)}{p(R|W) + p(W|R)} = p_\infty(R)$, say, as an upper limit. When $p_1(W) = 1$ and therefore $p_1(R) = 0$, eqn. (x) reduces to

$$p_t(R) = p_\infty(R) \{1 - [p(R|R) - p(R|W)]^t\} = b(1 - d_2^t) \quad \text{say.} \quad (\text{xii})$$

If we like to regard t as a continuous variable instead of a succession of discrete numbers, then the more general equation (xi) can evidently be written in the form

$$y = b(1 - ce^{-kt}), \quad (\text{xiii})$$

where $k = \log_e d_2 = \log_e [p(R|R) - p(R|W)]$.

Substituting in eqn. (xi) the numerical values found above, we obtain for the present data the equation $p_t(R) = .953(1 - .846^t)$. The estimated values, given by this equation, for the several trials have been entered in Table I (4th row). The discrepancies suggest that in point of fact the transitional probabilities do not remain constant but increase very slightly during the later stages of the training: however, owing to the small size of the sample the discrepancies are entirely non-significant.

Example 2. Most of our experiments have been carried out with children; and with these, as we shall see in a later paper, models that are somewhat more complex are generally required. When the process to be learnt is relatively simple, however, a reasonably good fit can often be obtained with the simple formulae already described. Once again, to show how close the estimates may become to the actual data, we select from our records an experiment in which the empirical frequencies show only a few and very small reversals of the general trend.

The apparatus used consisted of an upright wooden screen fixed to a flat horizontal base. In the screen were two holes, the one above the other, at which a light might appear. The screen concealed the manipulations of the experimenter at the back; and by means of two switch-keys he was able to decide at which hole the light should appear and what colour it should be—e.g., red, green, yellow or blue. On the base, within about equal reach of the child's right forefinger, were two buttons. The task of the child was to 'put out the light as quickly as possible by pressing the proper button.' For any one child the wiring was kept the same throughout the sitting. For other children the colours

¹ An excellent survey of the more important suggestions will be found in Gulliksen, *J. Gen. Psychol.* XI (1934), pp. 395 f.

The Statistical Analysis of the Learning Process

and the wiring were altered, so that no child who had already been tested could tell the others what to expect.¹

Fifty-eight educationally subnormal pupils aged 9 to 11 were tested. All had intelligence quotients between 60 and 70; and the mental age of the majority lay between 5 and 7. With most of them something like 25 to 30 trials were made. The number of correct responses during the first 20 trials (numbered 1 to 20) are shown in the top lines of Table IV.

TABLE IV. EXPERIMENTS ON TRAINING (2)

Correct Responses at Each Trial

Trial	1	2	3	4	5	6	7	8	9	10
Frequencies										
Observed	28	37	41	42	47	48	50	49	51	52
Estimated	29.0	36.7	40.4	43.9	46.4	48.2	49.5	50.6	51.2	51.7

Trial	11	12	13	14	15	16	17	18	19	20
Frequencies										
Observed	51	53	49	52	54	53	55	54	53	54
Estimated	52.0	52.3	52.5	52.6	52.7	52.8	52.8	52.9	52.9	52.9

The figures can be fitted tolerably well by the same simple method as before.² The transformation matrix (obtained by eqn. (x)) is $\begin{bmatrix} .976 & .253 \\ -.024 & .747 \end{bmatrix}$; and this gives the following formula for constructing the theoretical curve: $p_e(R) = .9136 + .4136(-.723)^R$. The figures so obtained are entered in the second line of Table IV. The agreement is unusually good, except that the formula appears to give an upper limit that is somewhat too low for the data. The excellence of the fit, however, is due to the fact that for illustrative purposes we deliberately picked a somewhat exceptional example.

Defects of the Method. There are, we believe, a few types of specifiable tasks for which this simple procedure is appropriate and adequate. For the most part they involve activities which, from the standpoint of the testee, are distinctly artificial. The process to be learned must be very elementary and very mechanical, and belong to a class of wider and more natural processes of which the testee has already had some small experience. In most of our experiments, however, the data obtained show a much poorer fit, and that, as a rule, in at least three respects. First, the empirical frequencies tend to vary above and below the smoothed curve in a far more irregular way: much of this zig-zagging no doubt is the result of the small samples available for testing; but some of the major undulations suggest the formation of temporary 'plateaux'. Secondly, in the latter half of the curve the frequencies often approach the limit much more rapidly than the formula would imply. Thirdly, and this is particularly true with complex tasks of a novel kind, the frequencies do not rise so rapidly as the formula implies during the earlier phases of the experiment.³

In such cases, except perhaps for purposes of a rough preliminary fit, we can no longer assume that the transitional probabilities remain constant. And it will then be difficult to construct a

¹ The apparatus was designed and constructed, in consultation with Miss Foley, by Mr. W. E. Ramsey. The original object of the investigation was, by the use of matched groups, to see whether children were likely to attend more readily to a difference of colour than to a difference of position, and how such tendencies changed with increasing age or intelligence.

² With this particular apparatus we originally supposed that the results could best be fitted with a complex formula derived *a priori* by taking into account the different options which (as the testee's observation ought quickly to discover) were at the experimenter's disposal: is the experimenter just as likely to be linking the difference of buttons with the difference of position as he is with the difference of colour? and in either case is he just as likely to link one button with this or that colour (or with this or that position) as he is to link it with the other? With a group of normal children a formula built up on these lines certainly seemed to give a slightly better fit. But with them the number of trials necessary were too few for anything like a definite curve of learning to be constructed; and with the most intelligent of all it was often possible to detect the moment at which the child suddenly formulated to himself the correct clues—after which all trials were correctly performed, except perhaps for a careless slip. With the feeble-minded children such a stage was never achieved.

³ Most forms of psychotherapeutic treatment (including remedial education) can be regarded as a form of training; and, when we plot an empirical curve to record the patient's improvement, a negative acceleration at the start, and occasional plateaux later on, prove to be common features. Since in the papers that follow we propose to regard such processes as essentially of the nature of learning, it is desirable to have a formula which will take due account of these peculiarities.

The validity of these and similar formulae might, we believe, be fruitfully tested by using data from electronic machines that claim to simulate human learning. Mr. G. Pask, of the Solartron Electronic Group, who demonstrated, at the recent Exhibition of the Physical Society, the performances of the machine he helped to design ('Eucrates I'), has promised his collaboration: the records of its progress in acquiring 'conditioned reflexes' and the like should furnish almost unlimited data obtained under precisely specified conditions.

suitable formula, unless we also abandon the assumption that the sequence is 'non-mnemonic' or (as statisticians sometimes say) 'non-hereditary'. That assumption, however, was merely a corollary to a general theory of causation. And, as we have already remarked, in dealing with psychological data, it is wiser to regard the earlier states of the system, not as the causes of the later states, but as the grounds for predicting or estimating them. In such a case there will be no reason for postulating any limited dependence of a strictly Markovian type. And accordingly we shall be justified in calculating autocorrelations between the sequence of frequencies as originally obtained and the same sequence after 2, 3, 4, or more displacements instead of after only one displacement. The factorization of a correlation (or product-sum) matrix of this kind requires factorial methods of a somewhat unusual type; and may yield, not just one significant and effective factor (as here),¹ but two or more. The identification of the factors and the deduction of appropriate formulae for fitting the curve throw considerable light on the nature of the psychological processes involved (as we shall hope to show in a subsequent paper); but separate discussions and special procedures are often required for each investigation.

Extensions of Formula. The type of equation to which such assumptions most frequently lead may be derived by a more direct argument. It will be seen that the simple formula we reached above (eqn. (xii)) implies that the rate of learning is proportional to the amount that still remains to be accomplished, i.e., $d/dt p_t(R) \propto 1 - p_t(R)$; it thus obeys a 'law of diminishing returns'. But the discrepancies that we have just noted indicate that a law of increasing returns is often more conspicuous during the early phases of the training. This suggests that the rate of learning is also proportional to the amount already achieved, i.e., $d/dt p_t(R) \propto p_t(R)$. Accordingly, combining both principles let us assume

$$dy/dt = k(a+y)(b-y) \quad (\text{xiv})$$

where a denotes the ability already possessed before the training starts, b the ability acquired when the training has achieved its utmost, and y the ability reached after the t th trial. Such an equation will yield a compound S-shaped curve, with increasing gains during the earlier phases and diminishing gains during the later.²

Inverting the derivative and integrating, we obtain

$$k(a+b) = \log_e(a+y)/(b-y) + C.$$

In the commonest type of record, $y = 0$ when $t = 0$. Evaluating the constant of integration on this basis, we obtain $C = -\log_e \frac{a}{b} = \log_e \frac{b}{a}$. Thus

$$k(a+b) = \log_e \frac{b(a+y)}{a(b-y)}, \text{ or } \frac{y}{b} = \frac{a(e^{k(a+b)t} - 1)}{a(e^{k(a+b)t} + b)}. \quad (\text{xv})$$

We can simplify this expression if we measure both the varying amount of progress and its maximum from an absolute zero, putting $Y = y + a$, $B = b + a$, and $C = b/a$. Then, on making the necessary substitutions ($y = Y - a$, $b = B - a$) in eqn. (xv), we obtain

$$Y/B = 1/(1 + Ce^{-Bkt}) \quad (\text{xvi})$$

$$= 1 - Ce^{-Bkt} + C^2e^{-2Bkt} - C^3e^{-3Bkt} + \dots \quad (\text{xvii})$$

If the trainee's experience has already included a good deal of relevant practice before the formal training actually starts, then a will be fairly large as compared with the maximum, and C fairly small as compared with Bk . Consequently, the terms containing the higher powers may be neglected, and a tolerably close approximation will be given by

$$Y = B(1 - Ce^{-Bkt}).$$

This simpler equation is identical in form with that reached above (eqn. xii).

The curve described by eqn. (xvi) is asymptotic to the horizontal lines $Y/B = 0$ and $Y/B = 1$. On setting the second derivative equal to zero, we find that the point of inflection occurs when

¹ Formally there are here two 'factors'; but only one is 'effective', i.e., is actually used in reconstructing the curve.

² Except for slight differences in notation the equations here derived are similar to those already proposed by Burt for describing the course of both learning and maturation in an earlier publication (*The Backward Child*, 1937, Appendix on 'Curves of Growth', pp. 644-51, eqn. 19, 20 and 21). As there observed, the curve obtained—the so-called 'logistic' curve—closely resembles the ogival curve for the normal probability integral. The curve was originally derived by Verhulst to express 'the law governing the growth of populations' (*Corr. Math. et Phys.* X, 1839, pp. 113-21.) It has since been used to describe a number of different chemical and biological phenomena. Perhaps the simplest and most instructive are the chemical transformations of the kind studied by van't Hoff and others. For instance, while eqn. (xii) describes the progress of a simple unimolecular reaction, eqn. (xvi) describes that of an autocatalytic unimolecular reaction: (cf. J. W. Mellor, *Chemical Statics and Dynamics*, 1904, pp. 245 f. and ref.). Eqn. (xvi) also appears to furnish a better description of the progress of epidemics than the equations of Pólya: here it can be assumed that the chances of infection are in proportion (i) to the number already infected and (ii) to the number of susceptible persons not yet infected who remain in the population. Then, eqn. (xvi) will give the total number of infected persons at time t , and eqn. (xviii) the rate at which new cases develop during different phases of the epidemic.

The Statistical Analysis of the Learning Process

$Y = \frac{1}{2}B$ or $y = \frac{1}{2}(b-a)$, that is, when $t = (1/Bk)\log_e\{(B-A)/A\} = t_0$ (say). Evidently the curve will be symmetrical about this point. Moreover, the rate of progress will be given by

$$\frac{dY}{dt} = kY(B-Y) = kB^2 \frac{Ce^{-Bkt}}{(1+Ce^{-Bkt})^2} = \frac{1}{4}kB^2 \operatorname{sech}^2[\frac{1}{2}kB(t-t_0)]. \quad (\text{xviii})$$

The curve so described is a rounded triangle or cocked hat, similar in shape to the normal curve of frequency and symmetrical about the ordinate through the point $(t-t_0)$. This suggests a plausible interpretation for the sigmoidal curve.¹ We may suppose (i) that the total process which the trainee is required to learn is made up of numerous constituent processes (or, in neurological terms, that it involves the formation of numerous neural connections), (ii) that the constituent items (or connections) vary widely in difficulty, and (iii) that the degree of difficulty is distributed roughly in accordance with the normal curve—i.e., that processes of medium difficulty are the commonest, and that the very easy constituents and the very difficult constituents are both comparatively rare. The consequences can readily be pictured if we think of some complex educational task, like learning a new language. The student masters the easiest elements first; but, since these are relatively few, his progress during the earlier stages will seem rather slow. As knowledge and skill increase, he is able to tackle the harder constituents; and, since these are far more numerous, his rate of improvement becomes more rapid. Towards the end, when only the hardest processes remain, he will proceed more slowly again, and the hardest items of all may remain for ever beyond his scope.

Eqn. (xvi) may easily be modified to deal with cases where the distribution of the difficulty is not symmetrical. The effect will be to modify the variable factor in the index, namely $(t-t_0)$, which we have hitherto supposed to have a direct or linear relation to the number of the trial or the duration of time, and substitute some more complex function of the number or the time. The function can then be expanded in a power series. This generalized version of the formula provides an excellent fit to curves that are not only asymmetrical but also undulating.

In fitting the constants required for a given set of data, it is usually easier to work with figures for the cumulative number of successes, S_t , say. Assuming that the trials are sufficiently numerous for us to treat the variables as continuous, we have, on integrating the expressions in eqn. (xvi),

$$S_t = B_t - \frac{B}{k} \{\log_e B - \log_e a + [(B-a)e^{-Bkt}]\} \quad (\text{xix})$$

Strictly speaking, the methods used in deducing eqn. (xiv) to (xix) are appropriate only to investigations where both N and n are large. For example, as one of us has endeavoured to show in an earlier publication (*The Backward Child*, 1937, p. 650), they will serve admirably to summarize results obtained with group testing and the like. Probably too they will suffice for the purpose of an elementary or introductory exposition. But in researches based on small samples more refined procedures are needed. Instead of differential equations it is better to employ linear difference equations; and instead of the deterministic form in which results like eqn. (xvi) are expressed, we

require the analogous stochastic formulae. If, for example, $\frac{\delta \bar{Y}}{\delta t}$ denotes the rate of change in $\bar{Y}$, we shall obtain

$$\frac{\delta \bar{Y}}{\delta t} = k(B\bar{Y} - \bar{Y}^2) \quad (\text{xx})$$

where $\bar{Y} = \mathcal{E}(Y)$, the mean or expected value of Y , and $\bar{Y}^2 = \mathcal{E}(Y^2)$. But the practical solution of the equations obtained by this procedure is decidedly laborious.²

Most of the simpler equations reached in this paper can be generalized to cover cases in which multiple responses are permitted instead of only two. And with these extensions, they include all the formulae that we shall apply in discussing the results of more elaborately planned investigations. In the later papers we propose to describe and analyse the results of further experiments and observations, chiefly on school children, which, we hope, may throw light on the factors underlying learning, remedial education, and psychotherapeutic treatment.

¹ Cf. *The Backward Child*, loc. cit. sup., pp. 646 f., where the relation between the above formulae and those of Thurstone and other investigators of the 'learning function' is more fully discussed.

² Cf. the treatment of the corresponding problem in the spread of epidemic disease in the papers by N. T. J. Bailey, *Biometrika*, XXXVII (1950), pp. 193 f. and *ibid.* XL (1953), pp. 177 f.

NOTES AND CORRESPONDENCE

WEIGHTED SUMMATION APPLIED TO PAIRED COMPARISONS

SIR,—Many teachers who are nowadays required to submit ratings of their pupils for various purposes (e.g. in connection with the examination for admission to grammar schools at the age of 11 plus) would be grateful for advice on a statistical problem which frequently confronts them. Recently three of my colleagues and I were asked to draw up an order of merit for six girls who fell within the requisite age-limits; and, though we could not agree on a general order, we found it possible in nearly every case to say which of any two was more intelligent than the other. We treated the results according to the method described by Professor Burt in an earlier issue of this *Journal* ('The Relative Merits of Ranks and Paired Comparisons', VI, 1953, Table III, p. 118), and counted up the total number of occasions on which each child was preferred to a fellow-pupil. The figures obtained are set out in Table I, and explain themselves: thus in the first column, the 1's show that Susan was unanimously rated more highly than Anne or Sylvia; the 0, that she was rated below Kate; and the $\frac{1}{2}$'s, that we could not agree over the last two girls: (in most cases this figure means that the two teachers who happened to know *both* the children to be compared failed to agree, while those who only knew one of them abstained from voting).

TABLE I. TEACHERS' JUDGEMENTS DERIVED BY COMPARING PAIRS

Number of times each girl named at the top of the column was judged more intelligent than the girl named at the left of each row.

Basis of Comparison		Children to be Judged						Total
		Susan	Anne	Sylvia	Kate	Mary	Jane	
Susan		($\frac{1}{2}$)	0	0	1	$\frac{1}{2}$	$\frac{1}{2}$	2 $\frac{1}{2}$
Anne		1	($\frac{1}{2}$)	0	0	$\frac{1}{2}$	$\frac{1}{2}$	2 $\frac{1}{2}$
Sylvia		1	1	($\frac{1}{2}$)	0	0	0	2 $\frac{1}{2}$
Kate		0	1	1	($\frac{1}{2}$)	0	1	3 $\frac{1}{2}$
Mary		$\frac{1}{2}$	$\frac{1}{2}$	1	1	($\frac{1}{2}$)	0	3 $\frac{1}{2}$
Jane		$\frac{1}{2}$	$\frac{1}{2}$	1	0	1	($\frac{1}{2}$)	3 $\frac{1}{2}$
Unweighted Total (0)		3 $\frac{1}{2}$	3 $\frac{1}{2}$	3 $\frac{1}{2}$	2 $\frac{1}{2}$	2 $\frac{1}{2}$	2 $\frac{1}{2}$	18
Weighted Total (1)		11 $\frac{3}{4}$	10 $\frac{3}{4}$	8 $\frac{1}{4}$	7 $\frac{1}{4}$	7 $\frac{1}{4}$	7 $\frac{1}{4}$	—
Weighted Total (2)		31 $\frac{5}{8}$	27 $\frac{5}{8}$	24 $\frac{7}{8}$	22 $\frac{5}{8}$	21 $\frac{5}{8}$	21 $\frac{5}{8}$	—
Proportions		100.0	87.4	78.7	71.6	66.8	68.4	—
...	...	...	...	...	...	...	...	...
Weighted Total (5):		...	...	...	...	...	...	...
Proportions		100.0	90.9	84.2	69.0	67.2	67.8	—
Percentages of Maximum		61.4	55.8	51.7	42.4	41.3	41.6	—

Note. The halves (in brackets) and the last five rows at the bottom have been inserted by the Editor for reasons explained below.

The totals (first row at the bottom of the table) appear to divide the pupils into two sub-groups—those who are brighter than the average (by half a mark only) and those who are duller (again by only half a mark). Somewhat to our surprise, there was no very clear distinction between the two groups, much less any definite order of merit. Nevertheless, at least two of us felt it possible to draw up such an order; and I have arranged the names at the head of the table in accordance with the order suggested by my own impressions.

Are we to conclude that our knowledge is too unreliable to warrant any such ranking? There are certainly several striking inconsistencies in our judgements. Yet they scarcely seem numerous enough to invalidate the judgements altogether. If, for example, one applies the test of 'homogeneity' or 'consistency' proposed by Burt (this *Journal*, VI, p. 9) we obtain $U = 0.39$, which is by no means a discouraging value.

Notes and Correspondence

To a teacher most of the apparent inconsistencies would seem to have quite an intelligible explanation. The first three girls are pupils from a higher class, though, being younger than their fellow-pupils, they take a low place there. The two teachers most familiar with that class agree in ranking them in the order indicated. Yet, although Susan is thus rated higher than Sylvia, and Sylvia than Kate, Kate (strange to say) is rated higher than Susan. Now Kate is in a lower class where she takes quite a high place; and the only subjects which the first three girls take in common with the lower class happen to be practical subjects which (at least in my own view) are comparatively easy (Needlework and Domestic Science). In these subjects Kate does much better than Susan. And there are similar explanations for the other inconsistencies.

Thus the evidence on which the children's relative ability is judged is slightly different with different pairs; with some it is primarily their prowess in academic subjects, with others their achievements in more practical subjects. Moreover, as every teacher would recognize, these latter subjects do not offer quite as trustworthy a basis for the assessments required. If so, should we not weight the various comparisons according to the difficulty of the subjects concerned? And, in that case, can the statistical psychologist suggest a suitable procedure? M. R. SEELY

Reply. The problem which Miss Seely has raised has been put by several correspondents, and therefore deserves discussion here. In its general character her table is analogous to the 'answer patterns' for a test of a dichotomous or trichotomous type (cf. this *Journal*, VI, pp. 7 f., Table I and refs.). Thus we may regard each row as designating a 'test', and the various entries as showing which children pass or fail in each 'test'. Accordingly, the totals can be taken as measuring the relative difficulty of the tests: comparison with Sylvia, for example, plainly forms a severer test than comparison with Kate, since only two children 'pass' the former, while three 'pass' the latter.

The chief peculiarity that distinguishes such tables from other 'answer patterns' is that they are necessarily square and always exhibit certain axisymmetric relations (any element below the diagonal can be obtained from the corresponding element above the diagonal by subtracting the latter from 1). Consequently, instead of using the row-totals to measure the ease of the several tests, we might, if we preferred, use the complementary column-totals to measure their difficulty.

As a matter of fact, much the same problems arise where the entries are marks obtained in *actual* tests—e.g., those used for the selection examination at the age Miss Seely has in mind. Ordinarily the marks would be added just as they stand. Frequently the totals then obtained for each test indicate wide differences in difficulty, particularly between the more academic subjects and the more practical. A similar difference seems discernible here; and in such cases (as Miss Seely suggests) the obvious device is to *weight* the marks with the totals, and then perform the addition afresh. But the new totals may then prove appreciably different from those that we took for our original weights. It would seem, therefore, as though we should repeat the process. Now it is not difficult to show that, if we keep on repeating our calculations, 'we shall by successive approximation eventually reach a stable set of factor measurements': factor analysis is indeed simply a short cut for reaching these ideal values; but 'the same factor measurements can be obtained by the method of weighted summation, applied direct to the observed marks'.¹

It may perhaps be convenient to give a formal proof. With the usual method of analysis and the usual notation, we should write

$$M = FP, \quad (1)$$

where M is a matrix of observed assessments or test-measurements (such as those in Table I above), P is the matrix of hypothetical factor-measurements, and F the matrix of factor-saturations (i.e., the hypothetical 'loadings' used in deducing the observed measurements from the factor-measurements). Here, however, M is necessarily a square matrix. In such a case we can always write

$$F = P^{-1}V, \quad (2)$$

$$\text{so that } M = P^{-1}VP, \quad (3)$$

where V is the diagonal matrix of latent roots.²

$$\text{Hence } PM = VP. \quad (4)$$

Now, if we assume that the best approximation to the true measurements for 'general intelligence' are given by the hypothetical measurements for the first and largest factor (the 'general factor'), then we need determine only the first row of P , say p'_1 . From equation (4) we have

$$p'_1 M = v_1 p'_1,$$

where v_1 denotes the first and largest latent root. Thus the row vector p'_1 contains the weights required. And the resulting totals, $v_1 p'_1$ (where v_1 is a scalar) will be proportional to the weights.

¹ Cf., C. Burt, 'The Influence of Differential Weighting', this *Journal*, III, pp. 114 f.; also *ibid.*, I, pp. 15-26 and ref.

² Cf., R. A. Frazer *et al.*, *Elementary Matrices*, 1938, p. 66, eqn. 11.

How then are we to determine this unknown set of weights? One obvious way of reaching this result would be to keep multiplying M by itself, and then take the weighted sum of these totals: for

$$M^n = P^{-1}VP \cdot P^{-1}VP \dots P^{-1}VP = P^{-1}V^nP;$$

and, since with empirical matrices the latent roots will be distinct, $P^{-1}V^nP$ will be approximately equal to $p_{1a}v^n p'_{1a}$, where p_{1a} denotes the first column vector in P^{-1} . By taking n large enough the approximation can be rendered as close as we wish.¹

In practice the simplest working procedure will be that which is commonly used in most other investigations where the same principle is involved: (see *Factors of the Mind*, Appendix I, 'Working Methods', pp. 467 f.). Instead of multiplying the matrix of marks by itself, it may be first pre-multiplied by a trial-set of weights; and the resulting product can be taken as giving an improved set of weights; the procedure must then be repeated until no essential difference remains between the proportional values. As usual, it is easiest to start with 1, 1, ..., 1 for the first trial-set; this means that we simply add each column as it stands (Table I: it is necessary, however, to include the diagonal figures for the comparison of each child with herself; the appropriate value is obviously $\frac{1}{2}$, which I have ventured to insert). Using the totals so obtained as a new set of weights, we get the next row of totals, which (to save space) I have added underneath Miss Seely's totals. As we are concerned with an agreement in the *proportionate* values of the weighted sums, not in their absolute values, it is simplest to reduce the weights first of all to fractions or percentages of the largest (cf. *Factors of the Mind*, loc. cit., Table VI). Continuing in this way we obtain the last set of weighted totals (marked 5), after which no appreciable change appears. These figures can be converted either to standard measure or to percentages of the maximum amount possible: here I have adopted the latter as being most intelligible to the teacher (last line of Table I).

The final figures now give a different mark to each of the six girls; and the children can therefore be ranked in order of merit. There is, it is true, but little difference between the last three. But the others are well spaced out. The order agrees with that suggested by Miss Seely, except that Jane now appears slightly better than Mary.

I should like to make one further suggestion. To measure the internal consistency of a table like Table I, which permits of *three* kinds of entry, 0, 1, and $\frac{1}{2}$, I suggest the following amplification of the formula given in my previous paper. Like that, it is based, as will be obvious, upon an analysis of the total variance of the marks into the variance due to the intercorrelation of the persons used as 'tests' and the variance expected in the absence of any such intercorrelation, namely,

$$U = \frac{3(\sum s_i^2 - n^3)}{n^3 - n},$$

where s_i denotes the unweighted sum for the i th column and n the number of persons.

It appears from Miss Seely's account that four teachers took part in the rating. If so, I think the results might be slightly more accurate if each teacher tabulated his own preferences separately (using $\frac{1}{2}$ when in doubt), and if the final table were obtained by summing their awards.

CYRIL BURT

¹ For the effect of thus 'powering' a square matrix, see C. Burt, 'The Unit Hierarchy and its Properties', *Psychometrika*, III, 1930, pp. 151 f. (esp. eqn. 29) and ref. A brief but clear explanation of the general procedure is given by Thomson in his *Factorial Analysis of Human Ability*, 1946, p. 75. The essential principle is that, 'as the matrix is repeatedly squared, it becomes more and more nearly hierarchical'.

The statistical reader will find an instructive discussion of the theoretical aspects of the present problem in T. H. Wei, 'The Algebraic Foundations of Ranking Theory' (Cambridge University Thesis, unpublished), and M. G. Kendall, 'Further Contributions to the Theory of Paired Comparisons', *Biometrics*, II, 1955, pp. 43-62.

THE INTERNATIONAL COLLOQUIUM ON FACTOR ANALYSIS

Paris, 11-16 July, 1955

The colloquium on factor analysis, arranged at Paris by the National Centre for Scientific Research, was well timed: the organizers, Professor Laugier and M. Reuchlin, are to be warmly commended for the admirable way in which the sessions were arranged; and the Rockefeller Foundation deserves the sincerest thanks of all who are interested in the progress of statistical psychology for the generosity with which it defrayed the expenses of the invited guests.

Of the nineteen 'participants' who had been invited to contribute papers nine were French: those from abroad included three pioneers in factor-analysis—Burt, Thurstone, and Hotelling; and the remainder came from Great Britain, the United States, Sweden, Germany, Spain, or Israel. The written contributions, which had been circulated in advance, were taken as read, and the main business of the meeting consisted in a debate on the issues raised. The programme has already been published in this *Journal* (VIII, p. 58). Since both papers and a summary of the discussions will be printed in the volume of proceedings, it will here be sufficient to mention points of special interest to British readers.

At the opening session Burt's introductory paper gave a 'historical survey of methods and results'; Thurstone's reviewed 'present problems and new methods of factorial analysis'; and Piéron's examined 'the general problem of the nature of factors in psychophysiology'. During the rest of the conference five main topics came up for discussion—the psychological characteristics of mental factors, the methodology and algebra of factorial techniques, the problems of statistical implication, the applications of factor analysis to psychological and biological issues, and the techniques of external multivariate analysis. The only important question on which no contribution was presented was the method of maximum likelihood.

In the first group of topics an outstanding contribution was that of J. P. Guilford; his dual subject was the dimensions of intellect and the factors involved in higher level thinking; he drew special attention to factors not usually considered in connection with intelligence testing, such as 'discovery', 'production', 'evaluation', and 'divergence-thinking'. Piéron reported a number of conclusions suggested by the results of 150 factor analyses carried out during the last ten years, and touched on such fundamental issues as the comparison of different methods and the value and significance of the factorial patterns disclosed. Yela concentrated on the psychological meaning of factors and factor analysis, and concluded that 'a certain kind of hierarchical multiple factor theory seems to be the most extended view among factorists today'. Madame Bernier's paper dealt with 'the number, identification and nature of psychological factors': she reviewed first the early contributions of Spearman, Burt's method of subdivided factors, and his technique of group factor analysis—all of which assume the presence of a general factor, and secondly Thurstone's concept of simple structure—which dispenses with any general factor, and incidentally stressed the advantages of Delaporte's ingenious *abaci* for testing significance, which had been but little used by psychologists. Reuchlin compared the factorial techniques developed by Burt and Thurstone respectively: Thurstone's rotated factors he equated with 'observed' factors and Burt's with 'theoretical'; he pointed out, however, that Burt's supplementary method of group factor analysis enabled 'observed factors' to be fitted into a scheme of hierarchical classification, such as that adopted by Moursy and British contributors.

Three papers were devoted solely to statistical issues. Darmais examined the probability relations between two groups of variables, and emphasized that general factor analysis appears practicable only when variables (transformed if necessary) allow the problem to be reduced to linear form. Delaporte sought methods of avoiding those cases of indeterminacy that often arise when traits are assumed to consist of a linear combination of independent common factors. Bargman presented a critical discussion of 'simple structure', and stressed the difficulty of deciding how many zero loadings should be postulated to define a statistically significant simple structure.

Papers on factorial methodology included a suggestion from Pineau for reducing the 'laborious computation often required by Hotelling's iterative methods of factor analysis'. Faverge sought to apply Spearman's scheme to the calculation of the 'images' of a criterion in relation to a given set of variables. Guttman reviewed the novel techniques which he has been developing during the last few years—many of them already described in the pages of this *Journal*, and explained in an illuminating way the essence of his *simplex*, *circumplex* and *radex* techniques.

Five papers dealt with applications. El Koussy illustrated the application of factor analysis to research on spatial abilities, and Huson and Henrysson its application to tests of achievement and the analysis of the school curriculum. Ledermann demonstrated its value for the study of mortality, and Schreider for the analysis of biological qualities: as was noted during the discussion, the results independently obtained by Schreider appeared to provide an interesting and an independent confirma-

tion of those obtained by Burt and Banks in their work on body measurements. Lastly, Peel took up the problem of factorizing correlations between persons and demonstrated how either external factor analysis or some variation of internal analysis may be used to identify the factors underlying a set of such correlations: the technique consisted in correlating the preferences exhibited by different persons in ranking a series of pictures or other objects with independent determiners describing the artistic qualities of these objects or pictures.

Several contributions were concerned with the relations of factorial analysis to various forms of external analysis. Hotelling's paper set forth in a lucid way the potentialities and limitations of both procedures. Peel's paper on aesthetic preferences provided concrete material which demonstrated the value of the external method of regression analysis. On the other hand, Eysenck contended that judging the validity of a battery of tests by an outside criterion might not always be appropriate. The view sponsored theoretically by Hotelling and illustrated by Peel, in contrast to that advocated by Eysenck, reflects a difference between two fields of interest: the educationist is concerned mainly with the prediction of external activity and so makes use of external criteria; the psychologist, working in the sphere of pure research, turns rather to internal analysis in accordance with the principles defended by Eysenck and implicit in the hierarchical scheme of factors suggested by Burt.

The discussions were throughout both lively and instructive. For English and American members it was a surprise to realize the large amount of factorial work that is now being carried out in France. It may seem invidious to single out special names; but I recall with some affection the delightful way in which Guttman seemed able to assimilate many divergent results to his *radex* theory; and all were impressed by the pertinent criticisms made by Bargman throughout the session. With so wide a range of subjects it would be difficult to condense the final outcome into a few broad generalizations like those which emerged from the Uppsala symposium of 1952. But it seems clear that we are entering a new phase in the development and application of factorial techniques. Its main features would appear to be first an increase in statistical rigour, secondly an insistence on the close relation between internal and external factor analysis, and thirdly an extension of factorial techniques to fields outside the narrower psychological material for which it was originally developed and used.

E. A. PEEL

JOURNAL COMMITTEE

During the past academic year the Journal Committee of the *British Journal of Statistical Psychology* has held two meetings. At the first (24th June 1955) the chief problems discussed were the arrangements for printing and publishing the journal and the contents of the issue for November 1955. The Editor explained that the Council of the Society had decided to accept the recommendation of the Managing Editor (Dr. Maule) who had advised the choice of the Aberdeen University Press as printer; the Publishers for the Society were to be Dr. Maule (as the Society's Treasurer) and Dr. Summerfield (as the Society's Secretary). During the meeting the Publishers presented a memorandum, drawn up in consultation with the Managing Sub-editor (Dr. Whitfield), outlining the duties of the publishers and editors respectively; the memorandum was discussed and accepted with minor modifications. In the course of the discussion several members drew attention to the importance of prompt publication, to the various recommendations submitted by the Publications Committee, and to the suggestions for increasing circulation recently put forward by several leading American psychologists in a letter to the *Bulletin* (May, 1955, pp. 62-3). The Publishers undertook to implement these proposals so far as possible. They pointed out that the size of separate issues and the style of printing were fixed by the terms of the agreement with the printers as approved by Council, but suggested that a pool of waiting manuscripts should be formed so that articles of appropriate size could be inserted as required. The Editor stated that endeavours were being made to include a series of elementary expository articles on topics already suggested by members of the Journal Committee (they would appear in the section headed 'Notes') and to increase the number of papers with an experimental as distinct from a purely theoretical basis.

At the following meeting (21st January 1956) the Editor reported the results of an informal conference between the Publishers, Editors, and President of the Society, regarding arrangements for publications during 1956 and for subsequent years, and put forward various suggestions for reducing cost of production, improving the size and style of type, and ensuring more punctual delivery. These were discussed in detail, and the Committee drew up a list of recommendations to be submitted to the Publications Committee of the Society at its forthcoming meeting. The contents for the May issue were discussed and approved.

CHARLOTTE BANKS, *Hon. Secretary*

BOOK REVIEW

Rank Correlation Methods. By M. G. KENDALL, 2nd edition. London: Charles Griffin & Co. 1955. Pp. vii + 196. 36s.

PROFESSOR KENDALL'S monograph has for seven years been used both as a textbook and as a manual—a dual rôle made possible by the relegation of the more exacting mathematical proofs to alternate chapters. The second edition, which incorporates the results of recent research, represents a considerable expansion of the first.

The principal alterations are of both theoretical and practical interest. During the last few years the relation between the two principal coefficients of rank correlation has been more thoroughly explored, with somewhat surprising results. For example, populations exist for which Kendall's coefficient τ takes the value zero, while Spearman's coefficient ρ takes the value ± 0.5 . And, more generally, the values of the two coefficients in the population and in the sample are connected by certain inequalities stated in the text. With disparities of this magnitude between the two coefficients, both of which are limited to the range $(-1, +1)$, it is clear that they must in some sense be measuring different aspects of the population; and more work is required to elucidate this point. Fortunately, however, these obscurities do not apply to bivariate normal populations. Here, unless the parental correlation is high, the analogues of the coefficients in the population cannot differ very widely; and the sample values of the two coefficients are very closely correlated, at least for large samples, except when the parental correlation is nearly 1.

Apart from questions of this sort, which are in the main stream of rank correlation theory, the tendency in recent years has been to treat ranking methods as part of the more general theory of distribution-free procedures. The author therefore had to decide whether or not to give a full account of the numerous applications and extensions of ranking theory, e.g., the applications recently made to such diverse problems as testing for trend in a time-series, testing the homogeneity of two or more populations, and estimating and comparing the strengths of association in contingency tables. Except for the last of these topics, he has decided against this further extension. The remainder are discussed in a short final chapter. Here he also refers very briefly to two new developments in the theory of paired comparisons—a procedure of special interest to psychologists. Of these the first is the model for paired comparisons proposed by Bradley and Terry of the Virginian Polytechnic Institute, where it has been investigated and found useful in work on taste-testing; the second is the construction of an overall rank order from a set of inconsistent paired comparisons—a problem of some complexity which he has discussed in a recent paper in *Biometrics*.

Although, as a result of the growth of the subject, this second edition of *Rank Correlation Methods* is less self-contained than the first, it remains an indispensable book for all who are seriously interested in the theory or application of such procedures.

ALAN STUART